The international journal of science / 9 July 2020

nature

ROBO CHEMIST

An autonomous approach to experimental chemistry

Seeing red
Three nations prepare to launch their missions to Mars

Rock weathering
A potential route to large-scale removal of atmospheric CO_2

Coronavirus
Human antibody blocks infection by SARS-CoV-2

The international journal of science / 9 July 2020

nature

The world goes to Mars

How to reach another planet when a pandemic hobbles yours.

On 15 July 1965, humanity got its first close-up look at Mars when NASA's Mariner 4 spacecraft flew past the red planet, recording grainy images of a barren, cratered surface. They were the first glimpse of another planet as seen from space.

Almost exactly 55 years later, 3 long-awaited Mars missions are due to launch (see page 184). Amid a coronavirus pandemic and raging geopolitical tensions, the missions, from the United States, China and the United Arab Emirates (UAE), are a powerful symbol of how nations can transcend their Earthly woes as they seek to explore and understand other worlds.

In the decades since Mariner 4, NASA has sent 19 missions to Mars, 4 of which failed. Today, the agency has three active missions orbiting the planet and two robots that are carrying out experiments on its surface. The latest US mission, Perseverance, which lifts off on 30 July at the earliest, is meant to push this exploration to the next level. It will roll around an ancient river delta in the Jezero Crater, searching for signs of past life. More importantly, it will drill into Martian rocks and collect rock and dirt samples as it travels. The ambition is for a future mission to land at Jezero, retrieve these rock samples and return them to Earth. If this happens, it would be the first-ever sample return from Mars — something researchers can't wait to analyse.

China's plan is just as ambitious. Later this month, the China National Space Administration intends to launch an orbiter, lander and rover combination called Tianwen-1, or 'quest for heavenly truth'. Many details have not yet been revealed, possibly because of the risk of failure — China tried unsuccessfully to send an orbiter to Mars in 2011. But it has pulled off several recent impressive accomplishments in space, including a series of Moon missions that culminated last year in the first mission to the lunar far side. The time may be right for Beijing to succeed in reaching Mars.

And then there is Hope, a Mars orbiter to be launched by the six-year-old UAE Space Agency (see page 190) no earlier than 15 July. It is the first interplanetary attempt by any Arab nation. Much of the spacecraft technology has been developed in collaboration with former NASA mission engineers hired by the UAE Space Agency. But the science is being primarily driven by Emirati researchers: a young and vibrant team of explorers. Hope aims to build the biggest, most-detailed map of Martian weather produced so far.

All three missions, which are due to arrive at Mars next February, need to launch in the next few weeks while Earth and Mars are in the best positions in their orbits for a spacecraft to travel between them — an event that happens only once every 26 months. It's remarkable that the coronavirus pandemic did not derail their plans. There was to have been a fourth Mars mission this summer, but the European Space Agency postponed its launch to 2022, in part because of the pandemic. NASA had to deploy some of its own planes to fly engineers between California and Perseverance's launch site in Florida because commercial flights were grounded. Meanwhile, China and the UAE both scrambled to finish their missions as COVID-19 raged.

> "Equally remarkable is that the three missions are not competing with each other."

Equally remarkable is that the three missions are not competing with each other, even though some commentators are calling the present state of US–China relations a new cold war. Whereas the original cold war between the Soviet Union and the United States dominated both nations' space ambitions in earlier decades, today's space agencies have relatively more-cooperative relationships.

That said, although NASA and the UAE Space Agency plan to make data from their missions publicly available, China's data policy remains unclear. China has been rolling out tranches of data from its Moon missions — the third batch from its lunar far-side mission was released last month. It should join the others, and pledge to share data from its Mars mission too.

Whereas intergovernmental relationships on Earth look ever more fraught, researchers must keep trying to transcend geopolitical squabbles. That includes ensuring that international collaboration on these missions continues, and that data are quickly made publicly accessible.

If these three emissaries launch successfully in the coming weeks, then we wait. We wait for them to traverse hundreds of millions of kilometres through the frigid vacuum of space, piloting themselves by the occasional command relayed from Earth. Red Mars will appear bigger as blue Earth grows smaller. They will arrive early next year at an alien, yet strangely familiar, planet. So, too, will we.

Pulling carbon from the sky is necessary, but not sufficient

Carbon dioxide removal is becoming a serious proposition — but it is not a substitute for aggressive action to cut emissions.

Could spreading basalt dust on farmers' fields help to remove atmospheric carbon? A large multidisciplinary team of scientists is confident it could, and that doing so could boost crop yields and soil health at the same time.

In this issue, David Beerling, a biogeochemist at the

University of Sheffield, UK, and his colleagues explore a strategy to enhance rock weathering (D. J. Beerling *et al.* *Nature* **583**, 242–248; 2020).

This is a continuously occurring natural phenomenon in which carbon dioxide and water react with silicate rocks on Earth's surface. In the process, atmospheric CO_2 is converted into stable bicarbonates that dissolve and then flow into rivers and oceans. The idea of scaling up this process to remove carbon has been considered for some three decades. The team's results provide the most detailed analysis yet of the technical and economic potential of this approach – and some of the probable challenges, including gaining public acceptance.

The researchers modelled what would happen to atmospheric carbon if basalt dust was added to agricultural lands in the world's biggest economies, including Brazil, China, the European Union, India, Indonesia and the United States. According to their calculations, doing so would remove between 0.5 billion and 2 billion tonnes of CO_2 from the air each year. The upper limit is more than 5 times the annual emissions of the United Kingdom, and akin to offsetting emissions from around 500 coal-fired power plants.

The team is also carrying out field trials in four countries – the only such trials yet. The authors have told *Nature* that preliminary results suggest the theory is holding up. The application of 20 tonnes of basalt dust to a half-hectare UK plot boosted CO_2 removal by 40% compared with that seen on an untreated plot, and by 15% in another trial, which spread dust over oil-palm plantations in Malaysia. The early results also indicate that adding basalt boosted yields in these and other crops.

These are encouraging developments at a time when governments around the world are struggling to meet their climate commitments. The approach, if successful, could enable high-emitting countries such as the United States and China to remove some of the carbon they have pumped into the atmosphere in recent decades. Moreover, the machines that are required to spread basalt dust on fields already exist: farmers use them to treat soils with limestone.

Costing the Earth

But, like many promising technological fixes, spreading basalt dust across the world's agricultural fields could prove more complicated than it first seems. Researchers must answer a host of pressing questions about the economic costs and environmental impacts. And there are potential questions for regulators, too.

Tinkering with the geochemical cycle will inevitably alter ecosystems in soils, rivers and even oceans. Some of this might be beneficial: rock dust of the right variety could bolster desirable plant communities, for example. And the alkaline content that runs off to the oceans could, in theory, counteract acidification, helping to protect corals and other creatures that are threatened by rising atmospheric CO_2 levels. But we need to be confident that there are no harmful consequences to land and sea, and any potential effects would need to be monitored carefully.

> **"Like many promising technological fixes, it could prove more complicated than it first seems."**

Moreover, mining rock on industrial scales, pulverizing it and spreading the dust on crop fields will not be cheap. The current price of carbon on the European Union's emissions trading system is less than €28 (US$31) per tonne. By contrast, Beerling and his colleagues estimate that enhanced rock weathering will cost between $80 and $180 per tonne of CO_2. That said, such costs are in line with competing technologies that could be used to pull CO_2 out of the atmosphere. And although rock will need to be mined, the Sheffield team is rightly calling for an inventory of free, suitable waste rock from existing mining operations. This will bring costs down, increase carbon uptake and make more efficient use of mined materials.

Citizen science

The project team also studied how members of the public would react to such technologies (E. Cox *et al. Nature Clim. Change* https://doi.org/10.1038/s41558-020-0823-z; 2020). From research carried out in the United Kingdom and the United States, it is clear that CO_2-removal strategies could face scepticism. Respondents who took part in surveys and workshop discussions feared that they might take too long to develop, and expressed concern that the basalt dust could affect ocean ecology. Many also opposed the idea of such technologies becoming a substitute for tackling the root causes of climate change.

Concerns surrounding the ecological impacts could be allayed with appropriate government oversight. But there is no intergovernmental process that is considering the full suite of issues – including safety and ethics – that will need to be addressed if carbon-removal technologies are to be applied at significant scales. The Carnegie Council for Ethics in International Affairs, a think tank in New York City, is working to build awareness among governments about the issues they are likely to face if these technologies are applied, through the Carnegie Climate Governance Initiative. Much of the group's work has been focused on how to regulate technologies associated with the 'geoengineering' label, such as lofting aerosols into the stratosphere to reflect solar radiation back into space. Carbon removal, although less controversial, is just as important.

Beerling and his colleagues also deserve credit on this front. The University of Sheffield's Leverhulme Centre for Climate Change Mitigation is 4 years into a 10-year, £10-million (US$12.5-million) research programme that includes modelling and field trials, as well as laboratory studies and public-engagement research. But the centre cannot be expected to shoulder such a heavy responsibility alone. Other groups and funders need to step up.

With the dangers of climate change becoming more apparent each year, countries must continue to pursue the aggressive action that will be required to meet the goals of the 2015 Paris climate agreement. Carbon-removal technologies cannot be a substitute for such action. But it is becoming clear that if humanity is to limit global warming to 1.5–2 °C above pre-industrial levels, it must pursue every promising idea.

A personal take on science and society

World view

By Pratyusha Kalluri

Don’t ask if AI is good or fair, ask how it shifts power

Those who could be exploited by artificial intelligence should be shaping its projects.

Law enforcement, marketers, hospitals and other bodies apply artificial intelligence (AI) to decide on matters such as who is profiled as a criminal, who is likely to buy what product at what price, who gets medical treatment and who gets hired. These entities increasingly monitor and predict our behaviour, often motivated by power and profits.

It is not uncommon now for AI experts to ask whether an AI is ‘fair’ and ‘for good’. But ‘fair’ and ‘good’ are infinitely spacious words that any AI system can be squeezed into. The question to pose is a deeper one: how is AI shifting power?

From 12 July, thousands of researchers will meet virtually at the week-long International Conference on Machine Learning, one of the largest AI meetings in the world. Many researchers think that AI is neutral and often beneficial, marred only by biased data drawn from an unfair society. In reality, an indifferent field serves the powerful.

In my view, those who work in AI need to elevate those who have been excluded from shaping it, and doing so will require them to restrict relationships with powerful institutions that benefit from monitoring people. Researchers should listen to, amplify, cite and collaborate with communities that have borne the brunt of surveillance: often women, people who are Black, Indigenous, LGBT+, poor or disabled. Conferences and research institutions should cede prominent time slots, spaces, funding and leadership roles to members of these communities. In addition, discussions of how research shifts power should be required and assessed in grant applications and publications.

A year ago, my colleagues and I created the Radical AI Network, building on the work of those who came before us. The group is inspired by Black feminist scholar Angela Davis’s observation that “radical simply means ‘grasping things at the root’”, and that the root problem is that power is distributed unevenly. Our network emphasizes listening to those who are marginalized and impacted by AI, and advocating for anti-oppressive technologies.

Consider an AI that is used to classify images. Experts train the system to find patterns in photographs, perhaps to identify someone’s gender or actions, or to find a matching face in a database of people. ‘Data subjects’ – by which I mean the people who are tracked, often without consent, as well as those who manually classify photographs to train the AI system, usually for meagre pay – are often both exploited and evaluated by the AI system.

Researchers in AI overwhelmingly focus on providing highly accurate information to decision makers. Remarkably little research focuses on serving data subjects. What’s needed are ways for these people to investigate AI, to contest it, to influence it or to even dismantle it. For example, the advocacy group Our Data Bodies is putting forward ways to protect personal data when interacting with US fair-housing and child-protection services. Such work gets little attention. Meanwhile, mainstream research is creating systems that are extraordinarily expensive to train, further empowering already powerful institutions, from Amazon, Google and Facebook to domestic surveillance and military programmes.

> **“It is our responsibility to recognize our skewed perspective and listen to those impacted by AI.”**

Many researchers have trouble seeing their intellectual work with AI as furthering inequity. Researchers such as me spend our days working on what are, to us, mathematically beautiful and useful systems, and hearing of AI success stories, such as winning Go championships or showing promise in detecting cancer. It is our responsibility to recognize our skewed perspective and listen to those impacted by AI.

Through the lens of power, it’s possible to see why accurate, generalizable and efficient AI systems are not good for everyone. In the hands of exploitative companies or oppressive law enforcement, a more accurate facial recognition system is harmful. Organizations have responded with pledges to design ‘fair’ and ‘transparent’ systems, but fair and transparent according to whom? These systems sometimes mitigate harm, but are controlled by powerful institutions with their own agendas. At best, they are unreliable; at worst, they masquerade as ‘ethics-washing’ technologies that still perpetuate inequity.

Already, some researchers are exposing hidden limitations and failures of systems. They braid their research findings with advocacy for AI regulation. Their work includes critiquing inadequate technological ‘fixes’. Other researchers are explaining to the public how natural resources, data and human labour are extracted to create AI.

Race-and-technology scholar Ruha Benjamin at Princeton University in New Jersey has encouraged us to “remember to imagine and craft the worlds you cannot live without, just as you dismantle the ones you cannot live within”. In this vein, it is time to put marginalized and impacted communities at the centre of AI research – their needs, knowledge and dreams should guide development. This year, for example, my colleagues and I held a workshop for diverse attendees to share dreams for the AI future we desire. We described AI that is faithful to the needs of data subjects and allows them to opt out freely.

When the field of AI believes it is neutral, it both fails to notice biased data and builds systems that sanctify the status quo and advance the interests of the powerful. What is needed is a field that exposes and critiques systems that concentrate power, while co-creating new systems with impacted communities: AI by and for the people.

Pratyusha Kalluri is a co-creator of the Radical AI Network and an AI researcher at Stanford University in California.
e-mail: pkalluri@stanford.edu

PRATYUSHA KALLURI

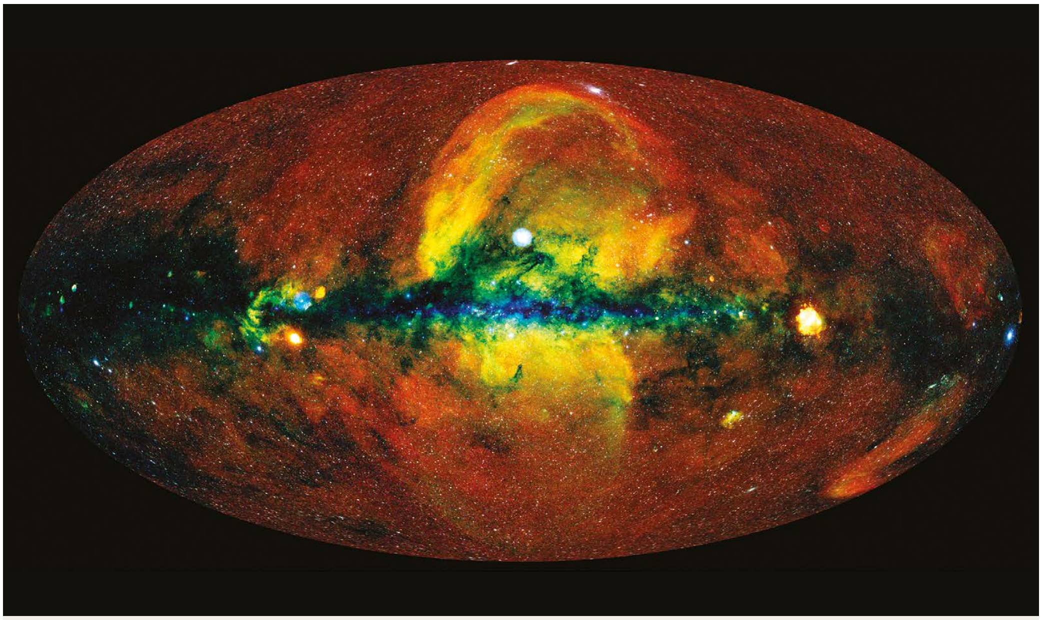
News in brief

CORONAVIRUS TEST FREQUENCY MATTERS MORE THAN SENSITIVITY

Communities such as universities, where COVID-19 cases could quickly spiral out of control, should test large numbers of people frequently for the new coronavirus – even if that means using a relatively insensitive test.

That's the conclusion of Michael Mina at the Harvard T.H. Chan School of Public Health in Boston, Massachusetts, and his colleagues, who modelled the effect of widespread testing on viral spread in a large group of people to gauge the importance of test sensitivity (D. B. Larremore *et al.* Preprint at medRxiv http://doi.org/d2gt; 2020). Tests that rely on the technique quantitative polymerase chain reaction (qPCR) can detect the merest traces of SARS-CoV-2 genetic material but are expensive and slow to return results.

The researchers found that weekly surveillance testing, paired with isolation of infected people, would limit an outbreak even if the testing method was less sensitive than qPCR. By contrast, surveillance testing done every 14 days would allow the total number of infections to climb almost as high as if there were no testing at all. The findings have not yet been peer reviewed.



JEREMY SANDERS, HERMANN BRUNNER & ESASS TEAM (MPE); EUGENE CHURAZOV, MARAT GILFANOV (ON BEHALF OF IKI)

Astronomers unveil epic X-ray map of the Universe

The newest map of the sky charted in high-energy X-rays offers a glimpse of what the Universe would look like if seen with X-ray vision. Researchers created the image using data from an instrument called eROSITA (Extended Roentgen Survey with an Imaging Telescope Array), part of the German–Russian satellite mission Spectrum-Roentgen-Gamma.

After sweeping the sky for six months, eROSITA has charted more than one million sources of X-ray radiation, including gigantic black holes, galactic clusters and the remnants of supernova explosions, many of which are new to science.

Researchers hope that a detailed map in this part of the spectrum will offer new ways to track the Universe's expansion, and to study the mysterious repulsive force called dark energy. The telescope will survey the X-ray sky for another 3.5 years.

"eROSITA has already revolutionized X-ray astronomy," said Kirpal Nandra, a high-energy physicist at the Max Planck Institute for Extraterrestrial Physics in Garching, Germany, where the project team is based. "But this is just a taste of what's to come."

ROGERIO PATEO/NAV/DAA UFMG

Fire tore through several rooms used to store plant, animal and human specimens at the Federal University of Minas Gerais in Brazil.

SECOND BRAZILIAN MUSEUM FIRE IN TWO YEARS REIGNITES CALLS FOR REFORM

Blaze at a natural history museum in Minas Gerais is forcing some researchers to relive the pain of losing priceless specimens and artefacts.

By Emiliano Rodríguez Mega

Researchers in Brazil are sifting through the ashes of a fire that destroyed part of a museum in the southeastern state of Minas Gerais on 15 June. The blaze follows repeated warnings about fire risks at museums, and comes less than two years after a massive inferno gutted the prized National Museum in Rio de Janeiro.

The latest fire has reopened wounds in the research community and intensified a national conversation about the need to protect Brazil's cultural and scientific heritage.

Mariana Lacerda, a geographer at the Federal University of Minas Gerais (UFMG) in Belo Horizonte, received a disturbing Monday-morning call: a building at the university's Natural History Museum and Botanical Garden, which she'd directed for almost a year, was in flames. When she arrived on the scene, smoke was still coming out of a single-storey building that housed thousands of artefacts, skeletal remains and taxidermied animals, many collected several decades ago.

Two storage rooms full of fossils and large archaeological objects were covered in soot and smoke. Flames had partly consumed a third room, which housed folk art, Indigenous artefacts and biological specimens. Two further rooms, housing important collections of insects, shells, birds, mammals, human bones and ancient plant remains, were almost completely lost.

For these, "little hope remains of material that can be recovered", Lacerda says. "Something that is so slow to build was destroyed so quickly, in just over an hour."

Archaeologist André Prous, who started

HISTORY IN FLAMES

Since 2010, at least six sites relevant to Brazil's scientific history have caught fire. There has also been damage to other cultural sites.

2010 — **Butantan Institute, São Paulo** — An inferno destroyed almost 90% of the museum's snake bank, the largest in Latin America, and a small part of its arachnid collection. Buntantan has been able to replace only about one-third of its snake specimens.

2012 — **Comandante Ferraz Antarctic Station** — A fire that started in the machine room housing the power generators destroyed approximately 70% of the research station. Two people died.

2013 — **Museum of Natural Sciences, Belo Horizonte** — The museum, owned by the Pontifical Catholic University of Minas Gerais, went up in flames in January.

2015 — **Museum of the Portuguese Language, São Paulo** — In December, a major fire destroyed the museum building and killed one firefighter. The museum has since been rebuilt and was scheduled to open its doors in June 2020.

2018 — **National Museum, Rio de Janeiro** — In September, a blaze at the National Museum claimed many of the most prized records of the nation's past. Recovery and reconstruction efforts continue.

2020 — **Natural History Museum and Botanical Garden, Belo Horizonte** — A fire destroyed parts of the museum owned by the Federal University of Minas Gerais, affecting stored zoological and archaeological specimens. As yet, authorities have no estimate of the scope of the loss.

working at the museum in 1975, was devastated. He and his colleagues had amassed a collection of human remains from a range of periods, including some from the earliest known inhabitants of Brazil, as well as samples of cultivated and wild plant species. Prous had also seen part of his life's work disappear during the 2018 fire at the National Museum, when ancient skulls that he helped to collect in the 1970s were destroyed.

"The sadness is matched only by the fear that other, similar disasters will continue to destroy [Brazil's] scientific heritage," he says. Some stone artefacts, ceramics and documentation of the sites he has excavated survived the blaze.

Historic losses

Brazilian museums have faced a series of fires, often resulting in irreparable losses, says Carolina Vilas Boas, director of museum processes at the Brazilian Institute of Museums in Brasilia. At least 12 buildings of cultural or scientific significance have burnt in the country, many of them in the past 10 years (see 'History in flames'). But the full extent of the damage is hard to know, says Vilas Boas, because reporting is probably incomplete.

Brazil is not unique in losing heritage institutions to fire, she says, but the country does have a poor record in taking care of its museums. Often, fire-prevention systems are installed, but budgets are too thin to maintain them properly. "There are many actions being taken to mitigate this risk," she says, but recurring economic crises have hindered long-term planning.

"That lack of resources had no relation to the fire in the collection's storage rooms," says Ricardo Hallal Fakury, a structural engineer at the UFMG. He did not speculate as to the cause of the fire, because investigations are still under way. But he says that the building that burnt was equipped with smoke detectors, and was mostly built of non-flammable materials.

Federal pressure

The tragedy in Belo Horizonte has amplified a decades-long discussion among Brazilian scientists pushing for national and state-level policies to help protect research collections, says Luciane Marinoni, an entomologist at the Federal University of Paraná and president of the Brazilian Society of Zoology, both in Curitiba. "The community is upset because we have been trying to solve this problem with the federal government but without success."

Some protective policies already exist. In 2017, the southern state of Paraná established norms and guidelines for the recognition of biological collections, defining who has responsibility for them, and putting in place objectives and goals to expand them and provide maintenance. Last year, the policy helped researchers to convince the government of Paraná to allocate 2 million reais (US$370,000) for the state's collections over the next three years. It's not a lot of money, but it's a solid start, says Marinoni: "The collections are leaving the darkness."

Back in Belo Horizonte, scientists are cleaning up after the fire. This time, however, they have some guidance on how to move forward.

National Museum researchers have teamed up with Lacerda to advise on the recovery of items that might still be salvageable. They are sharing protocols they developed after the 2018 blaze with UFMG professors and students who have volunteered to help. "Unfortunately, we are now experts in this matter," says palaeontologist Alexander Kellner, director of the National Museum. "We went through it. We know the mistakes to avoid, we have a way to act, we have a methodology."

PHYSICISTS FIND BEST EVIDENCE YET FOR ELUSIVE 2D STRUCTURES

Strange quasiparticles called anyons could herald a way to build quantum computers.

By Davide Castelvecchi

Physicists have reported what could be the first incontrovertible evidence for the existence of unusual particle-like objects called anyons, which were first proposed more than 40 years ago. Anyons are the latest addition to a growing family of phenomena called quasiparticles, which are not elementary particles, but are instead collective excitations of many electrons in solid devices. Their discovery — made using a 2D electronic device — could represent the first steps towards making anyons the basis of future quantum computers.

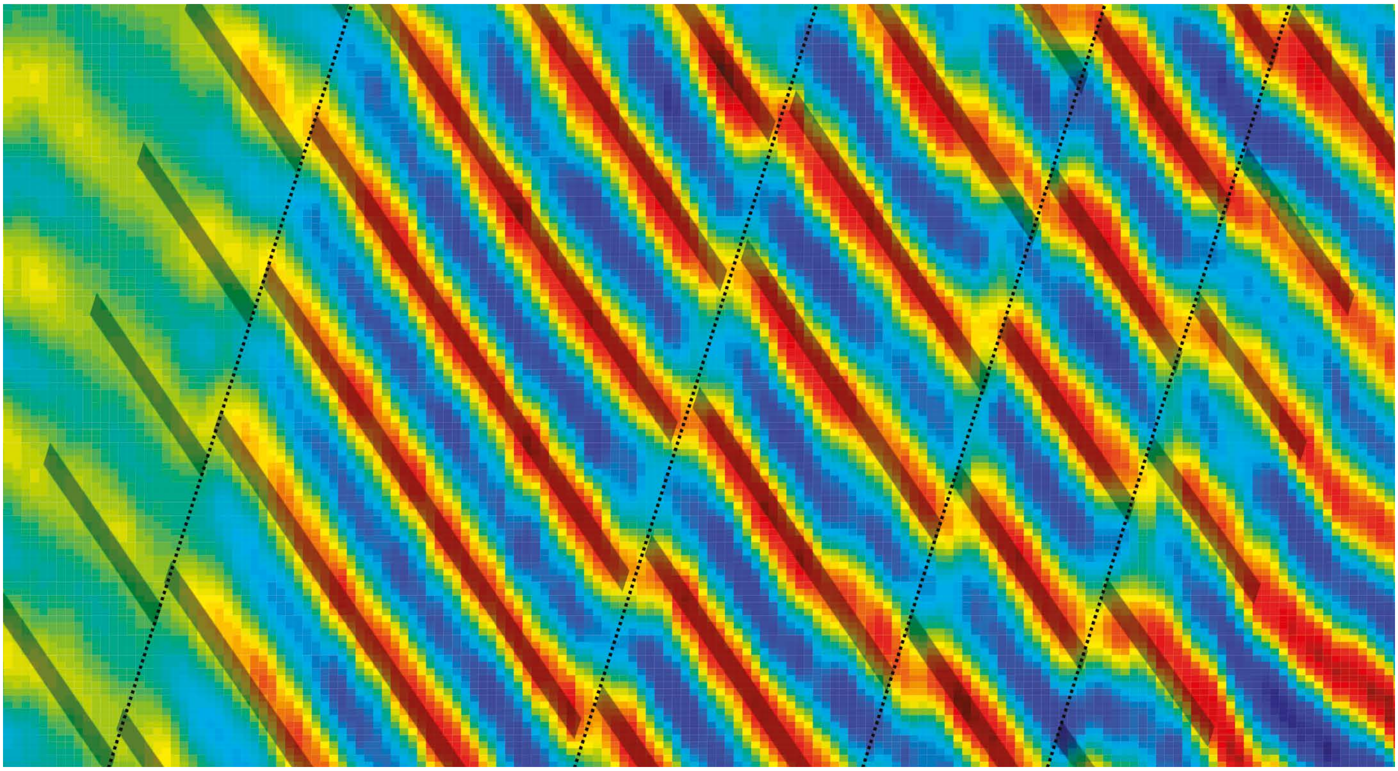
"This does look like a very big deal," says Steven Simon, a theoretical physicist at the University of Oxford, UK. The results, which have not yet been peer-reviewed, were posted on the arXiv preprint server last week[1].

Known quasiparticles display a range of exotic behaviours. For example, magnetic monopole quasiparticles have only one magnetic pole — unlike all ordinary magnets, which always have a north and a south. Another example is Majorana quasiparticles, which are their own antiparticles.

Anyons are even more strange. All elementary particles fall into one of two possible categories — fermions and bosons. Anyons are neither. The defining property of fermions (which include electrons) is Fermi statistics: when two identical fermions switch spatial positions, their quantum-mechanical wave — the wavefunction — is rotated by 180°. When bosons exchange places, their wavefunction doesn't change. Switching two anyons should produce a rotation by some intermediate angle. This effect, which is called fractional statistics, cannot occur in 3D space, but only as collective states of electrons confined to move in two dimensions.

Fractional statistics

Fractional statistics is the defining property of anyons, and the latest work — led by Michael Manfra, an experimental physicist at Purdue University in West Lafayette, Indiana — is the



JAMES NAKAMURA & MICHAEL MANFRA

The ‘pyjama stripe’ interference pattern denotes the presence of anyons in an electronic system.

first time it has been measured so conclusively.

The quasiparticles’ unusual behaviour when switching places means that if one particle moves in a full circle around another – equivalent to the two particles switching positions twice – it will retain a memory of that motion in its quantum state. That memory is one of the telltale signs of fractional statistics that experimentalists have been looking for.

Manfra and his team manufactured a structure consisting of thin layers of gallium arsenide and aluminium gallium arsenide. This confines electrons to move in two dimensions, while shielding them from stray electric charges in the rest of the device. The researchers then cooled it to 10,000ths of a degree above absolute zero and added a strong magnetic field. This produced a state of matter in the device called a fractional quantum Hall (FQH) insulator, which has the peculiarity that no electric current can run in the interior of the 2D device, but can run along the edge. FQH insulators can host quasiparticles whose electric charge is not a multiple of the electron charge, but is instead one-third of it: these quasiparticles have long been suspected to be anyons.

To prove that they had indeed detected anyons, the researchers etched the device so that it could carry currents from one electrode to another along two possible edge paths. The team tweaked the conditions by varying the magnetic field and adding an electric field. These tweaks were expected to create or destroy anyon states stuck in the interior, and also to produce anyons running between the electrodes. Because moving anyons had two possible paths, each producing a different twist in their quantum-mechanical waves, when the anyons reached the end point, their wavefunctions produced an interference pattern called pyjama stripes.

This pattern shows how the relative amount of rotation between the two paths varies in response to changes in the voltage and the magnetic-field strength. But the interference also displayed jumps, which were the smoking gun[2] for the appearance or disappearance of anyons in the bulk of the material.

“As far as I can tell, it is an extremely solid observation of anyons – directly observing their defining property: that they accumulate a fractional phase when one anyon travels around another,” Simon says.

It is not the first time that researchers have reported evidence of fractional statistics. Robert Willett, a physicist at Nokia Bell Labs in Murray Hill, New Jersey, says that his team saw “strong evidence” for fractional statistics in 2013 (ref. 3).

Quantum computing

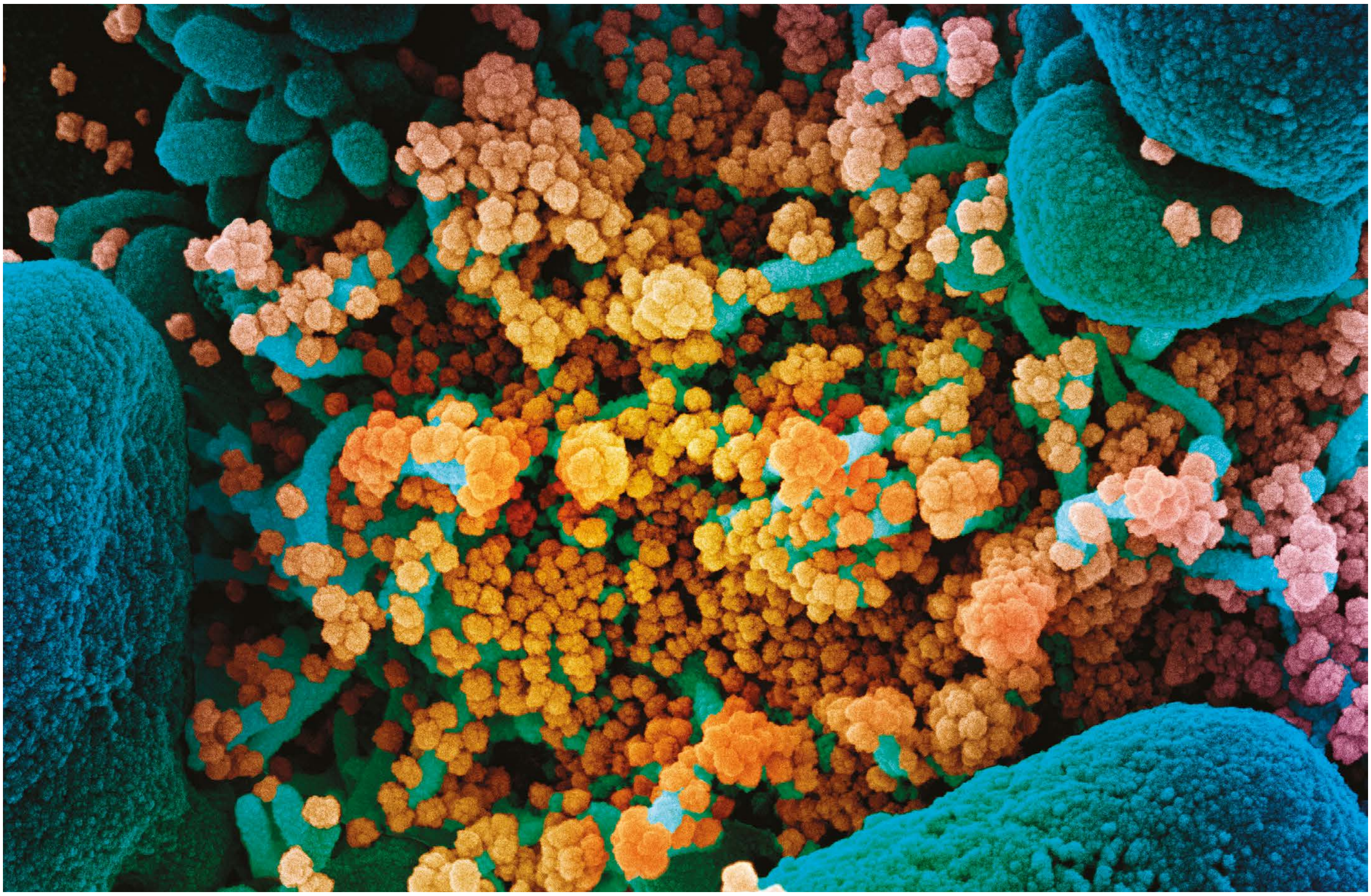
But some theoretical physicists say that the evidence in these and other experiments, although striking, was not conclusive. “In many cases, there are several ways of explaining an experiment,” says Bernd Rosenow, a condensed-matter theorist at the University of Leipzig in Germany. But the evidence reported by Manfra’s team, if confirmed, is unequivocal, Rosenow says. “I’m not aware of an explanation of this experiment which is plausible and does not involve fractional statistics.”

The results potentially lay the groundwork for applications for anyons. Simon and others have developed elaborate theories for how anyons could be used as the platform for quantum computers. Pairs of the quasiparticles could encode information in their memory of how they have circled around one another. And because the fractional statistics is ‘topological’ – it depends on the number of times one anyon went around another, and not on slight changes to its path – it is unaffected by tiny perturbations. This robustness could make topological quantum computers easier to scale up than are current quantum-computing technologies, which are error-prone.

Topological quantum computing will require more-sophisticated anyons than those Manfra and colleagues have demonstrated; his team is now redesigning its device to achieve that. Still, anyon applications are some way off, researchers warn. “Even with this new result, it is very hard to see [fractional quantum-Hall] anyons as a strong contender for quantum computing,” Simon says.

But the quasiparticles’ unique physics is worth exploring: “To me, as a condensed-matter theorist, they are at least as fascinating and exotic as the Higgs particle,” says Rosenow.

1. Nakamura, J., Liang, S., Gardner, G. C. & Manfra, M. J. Preprint at https://arxiv.org/abs/2006.14115 (2020).
2. Rosenow, B. & Stern, A. *Phys. Rev. Lett.* **124**, 106805 (2020).
3. Willett, R. L. *et al. Phys. Rev. Lett.* **111**, 186401 (2013).



A scanning electron microscope image of SARS-CoV-2 coronavirus particles (orange) on a cell (blue).

SIX MONTHS OF CORONAVIRUS: THE MYSTERIES SCIENTISTS ARE STILL RACING TO SOLVE

From immunity to the role of genetics, *Nature* looks at five pressing questions about COVID-19 that researchers are tackling.

By Ewen Callaway, Heidi Ledford and Smriti Mallapaty

In late December 2019, reports emerged of a mysterious pneumonia in Wuhan, China, a city of 11 million people in the province of Hubei. The cause, Chinese scientists quickly determined, was a new coronavirus distantly related to the SARS virus that had emerged in China in 2003, before spreading globally and killing nearly 800 people.

Six months and more than ten million confirmed cases later, the COVID-19 pandemic has become the worst public-health crisis in a century. More than 500,000 people have died. It has also catalysed a research revolution, as researchers and doctors have worked at breakneck speed to understand COVID-19 and the virus that causes it: SARS-CoV-2.

They have learnt how the virus enters and hijacks cells, how some people fight it off and how it eventually kills others. They have identified drugs that benefit the sickest patients, and many more potential treatments are in the works. And researchers have developed nearly 200 potential vaccines.

But for every insight into COVID-19, more questions emerge, and others linger. That is how science works. To mark six months since the world first learnt about the disease responsible for the pandemic, *Nature* runs through some key questions that researchers still don't have answers to.

Why do people respond so differently?

One of the most striking aspects of COVID-19 is the stark differences in experiences of the disease. Some people never develop symptoms, whereas others, some apparently healthy, have severe or fatal pneumonia. "The differences in the clinical outcome are dramatic," says Kári Stefánsson, a geneticist and chief executive of DeCODE Genetics in Reykjavik, which is looking for human gene variants that might explain some of these differences.

That search has been hampered by the small number of cases in Iceland. But last month, a team analysing the genomes of roughly 4,000 people from Italy and Spain turned up the first strong genetic links to severe COVID-19 (ref. 1). People who developed respiratory failure were more likely to carry one of two particular gene variants than were people without the disease.

One variant lies in the region of the genome that determines ABO blood type. The other is near several genes, including one that encodes a protein that interacts with the receptor the virus uses to enter human cells, and two others that encode molecules linked to immune response against pathogens. The researchers are part of the COVID-19 Host Genetics Initiative, a global consortium of groups that are pooling data to validate findings and uncover further genetic links.

The variants identified so far seem to play a modest part in disease outcome. A team led by Jean-Laurent Casanova, an immunologist at the Rockefeller University in New York City, is looking for mutations that have a more substantial role.

What's the nature of immunity and how long does it last?

Immunologists are working feverishly to determine what immunity to SARS-CoV-2 could look like, and how long it might last. Much of the effort has focused on 'neutralizing antibodies', which bind to viral proteins and directly prevent infection. Studies have found[2] that levels of neutralizing antibodies against SARS-CoV-2 remain high for a few weeks after infection, but then typically begin to wane.

However, these antibodies might linger at high levels for longer in people who had particularly severe infections. "The more virus, the more antibodies, and the longer they will last," says immunologist George Kassiotis of the Francis Crick Institute in London. Similar patterns were seen with SARS (severe acute respiratory syndrome).

Researchers don't yet know what level of neutralizing antibodies is needed to fight off reinfection by SARS-CoV-2. And, ultimately, a full picture of SARS-CoV-2 immunity is likely to extend beyond antibodies. Other immune cells called T cells are important for long-term immunity, and studies[3,4] suggest that they are also being called to arms by SARS-CoV-2.

Has the virus developed any worrying mutations?

All viruses mutate as they infect people, and SARS-CoV-2 is no exception. Molecular epidemiologists have used these mutations to trace the global spread of the virus. But scientists are also looking for changes that affect its properties, for instance by making some lineages more or less virulent or transmissible. "If it did become more severe, that's something you would want to know about," says David Robertson, a computational biologist at the University of Glasgow, UK, whose team is cataloguing SARS-CoV-2 mutations. Such mutations also have the potential to lessen the effectiveness of vaccines, by altering the ability of antibodies and T cells to recognize the pathogen.

But most mutations will have no impact, and picking out the worrying ones is challenging. Versions of the coronavirus identified at the start of outbreaks in hotspots such as Lombardy in Italy or in Madrid, for instance, might look as if they are deadlier than those found at later stages or in other locations. But such associations are probably spurious, says William Hanage, an epidemiologist at Harvard University's T.H. Chan School of Public Health in Boston, Massachusetts: health officials are more likely to identify severe cases in early, uncontrolled stages of an outbreak. Broad spread of certain mutations could also be due to 'founder effects', in which lineages that arise early in transmission centres such as Wuhan or Italy happen to have a mutation that is passed on when they seed outbreaks elsewhere.

Horseshoe bats might be the origin of the virus.

MENAHEM KAHANA/AFP/GETTY

Researchers are debating whether the widespread prevalence of one mutation in the virus's spike protein is the product of a founder effect — or an example of a consequential change to the virus's biology. The mutation seems to have first emerged around February in Europe, where most circulating viruses now carry it, and it is currently found in every region of the world. Studies have suggested that this mutation makes the SARS-CoV-2 virus more infectious to cultured cells, but it is not clear how this translates to human infections.

How well will a vaccine work?

An effective vaccine might be the only way out of the pandemic. There are currently roughly 200 in development. The first large-scale efficacy trials to find out whether any vaccines work are set to begin in the next few months. These studies will compare rates of COVID-19 infection between people who get a vaccine and those who receive a placebo.

But there are already clues in data from animal studies and early-stage human trials, mainly testing safety. Multiple teams have conducted 'challenge trials' in which animals given a candidate vaccine are intentionally exposed to SARS-CoV-2 to see whether the jab can prevent infection. Studies in macaques suggest that vaccines might prevent lung infection and resulting pneumonia, but not block infection elsewhere in the body, such as the nose. Monkeys that received a vaccine developed by the University of Oxford, UK, and that were then exposed to the virus had levels of viral genetic material in their noses comparable to levels in unvaccinated animals[5]. Results such as this raise the possibility of a COVID-19 vaccine that prevents severe disease — but not spread of the virus.

Data in humans, although scant, suggest that COVID-19 vaccines prompt our bodies to make potent neutralizing antibodies that can block the virus from infecting cells. What isn't yet clear is whether levels of these antibodies are high enough to stop new infections, or how long these molecules persist in the body.

What is the origin of the virus?

Most researchers agree that the SARS-CoV-2 coronavirus probably originated in bats, specifically horseshoe bats. This group hosts two coronaviruses closely related to SARS-CoV-2. One, named RATG13, was found[6] in intermediate horseshoe bats (*Rhinolophus affinis*) in the southwestern Chinese province of Yunnan in 2013. Its genome is 96% identical to that of SARS-CoV-2. The next-closest match is RmYN02, a coronavirus found in Malayan horseshoe bats (*Rhinolophus malayanus*), which shares 93% of its genetic sequence with SARS-CoV-2 (ref. 7).

The 4% difference between the genomes of RATG13 and SARS-CoV-2 represents decades of evolution. Researchers say this suggests that the virus might have passed through an intermediate host before spreading to people, in the same way that the virus that causes SARS is thought to have passed from horseshoe bats to civets before reaching people.

To unequivocally trace the virus's journey to people, scientists would need to find an animal that hosts a version more than 99% similar to SARS-CoV-2 — a prospect complicated by the fact that the virus has spread so widely among people, who have also passed it to other animals, such as cats, dogs and farmed mink.

Zhang Zhigang, an evolutionary microbiologist at Yunnan University in Kunming, says efforts by research groups in China to isolate the virus from livestock and wildlife, including civets, have turned up bare. Groups are also searching for the coronavirus in tissue samples from bats, pangolins and civets.

1. Ellinghaus, D. *et al*. Preprint at medRxiv https://doi.org/10.1101/2020.05.31.20114991 (2020).
2. Long, Q.-X. *et al*. *Nature Med*. https://doi.org/10.1038/s41591-020-0965-6 (2020).
3. Grifoni, A. *et al. Cell* **181**, 1489–1501 (2020).
4. Ni, L. *et al. Immunity* **52**, 971–977 (2020).
5. van Doremalen, N. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2020.05.13.093195 (2020).
6. Zhou, P. *et al. Nature* **579**, 270–273 (2020).
7. Zhou, H. *et al. Curr. Biol.* **30**, 2196–2203 (2020).

Q&A


JAYAPRAKASH MULIYIL

India's epidemic adviser fears coronavirus crisis will worsen

India has confirmed more than 700,000 cases of COVID-19 so far, making it the world's third-worst-hit country. Major cities such as Delhi and Mumbai are particularly badly affected, with hospitals struggling to accommodate critically ill patients. The current surge in infections follows a two-and-a-half-month India-wide lockdown that began on 25 March and severely disrupted the economy and livelihoods. Jayaprakash Muliyil, an epidemiologist at the Christian Medical College in Vellore in the state of Tamil Nadu, has been advising the Indian government on COVID-19 surveillance. He talks to *Nature* about some of the factors affecting India's epidemic.

Do you think the outbreak in India is charting a different path from outbreaks in other badly hit countries, such as the United States, Italy or Spain?
It is. It is spreading much faster here, and the infection rates are higher. The general population's anxiety about the disease is low. People will willingly go out into the market today, and take fewer precautions to protect themselves. Consequently, at least in cities, the epidemic is growing very rapidly. And we know it is spreading in rural areas, too. The whole trajectory of the infection is moving upwards more sharply than in many other countries. What happened in many Western countries is that when a big city like London was affected, other cities reacted strongly and reduced transmission. So, everywhere else, the doubling time got longer, but in some Indian cities it is short.

India is reporting that its mortality rate is among the lowest in the world. Is that accurate?
The mortality per million people in India is expected to be lower because of the low average age of India's population. (Older people are more likely to die from this infection.) So, we can take some comfort in the fact that deaths are fewer, especially in the rural population.

But the problem with death as an indicator is that a COVID-19 death has to be certified as such. The only way to do this is through an RT-PCR test (a reverse-transcription polymerase chain reaction test, which looks for viral genetic material in nose and throat samples). And with a population of 1.3 billion, what do you think is the proportion of people that has access to this kind of testing? It is very low.

So, it is very difficult to count all deaths due to COVID-19. There is no way you can get it done, unless rapid tests become more widely available. Remember that at least half of all deaths will happen in the rural villages where around 66% of our population live. And there are no real mechanisms to ascertain causes of deaths in these villages.

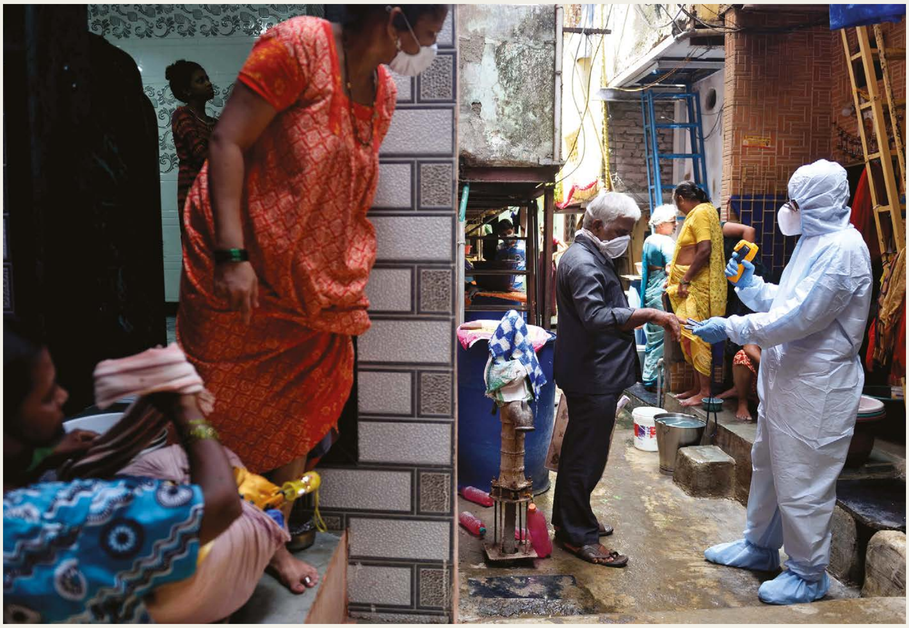
What do you think of the Indian government's response to the epidemic so far?
The lockdown all over the country was not the right response. It brought misery to untold numbers of people and destroyed lives. And we haven't been able to repair its consequences for society. That was unfortunate. If we had planned the lockdown better, we would have still had losses, but they wouldn't have been greater than what we are experiencing now.

The lockdown did have one benefit: everyone became aware of this thing called COVID-19. It is not easy to communicate this to everyone in India, with its many remote regions, but because of the lockdown, people heard about it. The concept of an infectious disease is not easy for many to understand. In many rural areas, measles is considered to be caused by a goddess visiting a village. So is chicken pox. There, when you introduce the term virus, it doesn't make sense to many groups of people.

What should cities with large outbreaks do now?
Many cities are quarantining people returning from COVID-19-affected states or countries in public facilities and hotels. I would say that should stop, and these people should quarantine at home. Most of them won't know whether they have been infected, because they might not have been tested. And when the number of infected people is already high in the community, quarantining incoming travellers in public facilities, which is very labour-intensive, is not economical.

Instead, we should focus on a reverse quarantine for elderly people — where the old and the vulnerable are quarantined from others to protect them. We should also put money into hospitals, and provide oxygen for patients. That manoeuvre will save lives.

Interview by Priyanka Pulla
This interview has been edited for length and clarity.


Temperature screening in Mumbai, India, during the COVID-19 pandemic.

SATISH BATE/HINDUSTAN TIMES VIA GETTY

CANCER RESEARCHER'S APPOINTMENT PROMPTS MAJOR ROW

Pier Paolo Pandolfi, who harassed a postdoc while at Harvard, will no longer lead an Italian institute.

By Alison Abbott

An Italian scientific institute has reversed its decision to appoint a high-profile cancer researcher – Pier Paolo Pandolfi – as its scientific director after a tumultuous month of protests and accusations. The row over Pandolfi, who admits one instance of sexual harassment and has been accused of scientific misconduct in his research papers, resulted in the resignation of the entire scientific board of the Veneto Institute of Molecular Medicine (VIMM) in Padua.

Pandolfi was director of the cancer-genetics centre at Harvard University's prestigious Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, until last year, and has made discoveries about the molecular mechanisms and genetics underlying some cancers. Since May, reports that he had been accused of sexual harassment have appeared online. Meanwhile, over the past seven years, anonymous commenters on the website PubPeer, which hosts discussion of published research results, have raised questions about the integrity of some of Pandolfi's papers.

Pandolfi admits to the inappropriate pursuit of a postdoc at Harvard, for which he says Harvard investigated him. He says it was an isolated incident and he has received counselling, and he resigned from Harvard last December. He denies the accusations of research misconduct in his work, but says that he will review any papers under scrutiny.

Mass resignation

The foundation that funds VIMM chose Pandolfi as the institute's scientific director on 20 May. That decision sparked protests by the institute's entire scientific advisory board, which includes two Nobel prizewinners. The members resigned en masse on 25 June after the appointment was confirmed. However, they say that although they were aware of the allegations against Pandolfi, their mass resignation was protesting against the fact that they had not been consulted in the appointment procedure. Many of VIMM's principal investigators also said that they had not been consulted, despite the institute's statutes requiring this, and said that the concerns about Pandolfi should have been investigated. VIMM's interim scientific director, Luca Scorrano, resigned on 22 June over the situation.

Cancer researcher Pier Paolo Pandolfi.

JAMES DEREK DWYER/BETH ISRAEL DEACONESS MEDICAL CENTER (CC BY-SA 4.0)

Giving advice on the appointment of key scientific staff is a central role of such boards, says Wolfgang Baumeister, chair of VIMM's scientific advisory board, and a director of the Max Planck Institute of Biochemistry in Martinsried, Germany. He says it was "totally unacceptable" for the foundation to appoint Pandolfi without consulting the board.

> "There should have been more investigation before making the appointment."

"There should have been more investigation before making the appointment," says board member Aaron Ciechanover, a biochemist at the Technion – Israel Institute of Technology in Haifa who shared the 2004 chemistry Nobel prize.

Under pressure from the scientists and from the Italian media, which has reported on the allegations, the foundation reversed its decision on 30 June. In a statement, the foundation's directors said that rescinding the appointment became "necessary, after learning the story in which Prof. Pandolfi was involved in Harvard University, of which the Foundation had not been informed". Appointing him would compromise "the image and reputation of the Institute," they said.

The foundation's president and chair of its executive board, Francesco Pagano, a urologist who made the decision to appoint Pandolfi with the board, says that Pandolfi had not told him or the foundation about the allegations when Pagano made the appointment, and that he learnt about them from the press. Pagano says that, according to the institute's statutes, the scientific advisory board is only a consulting body for scientific questions, and that the principal investigators were consulted appropriately.

Frequent e-mails

When it comes to the harassment, Pandolfi says that the incident in which he pursued the postdoc was isolated. "It was romantic, not sexual – and it was the biggest mistake in my life," he says. He says that an internal Harvard investigation, which concluded in July, referred him to an external service for evaluation and coaching. Harvard declined to comment on whether it had investigated Pandolfi because it said it does not comment on personnel issues, but it confirmed that he is no longer affiliated with the BIDMC.

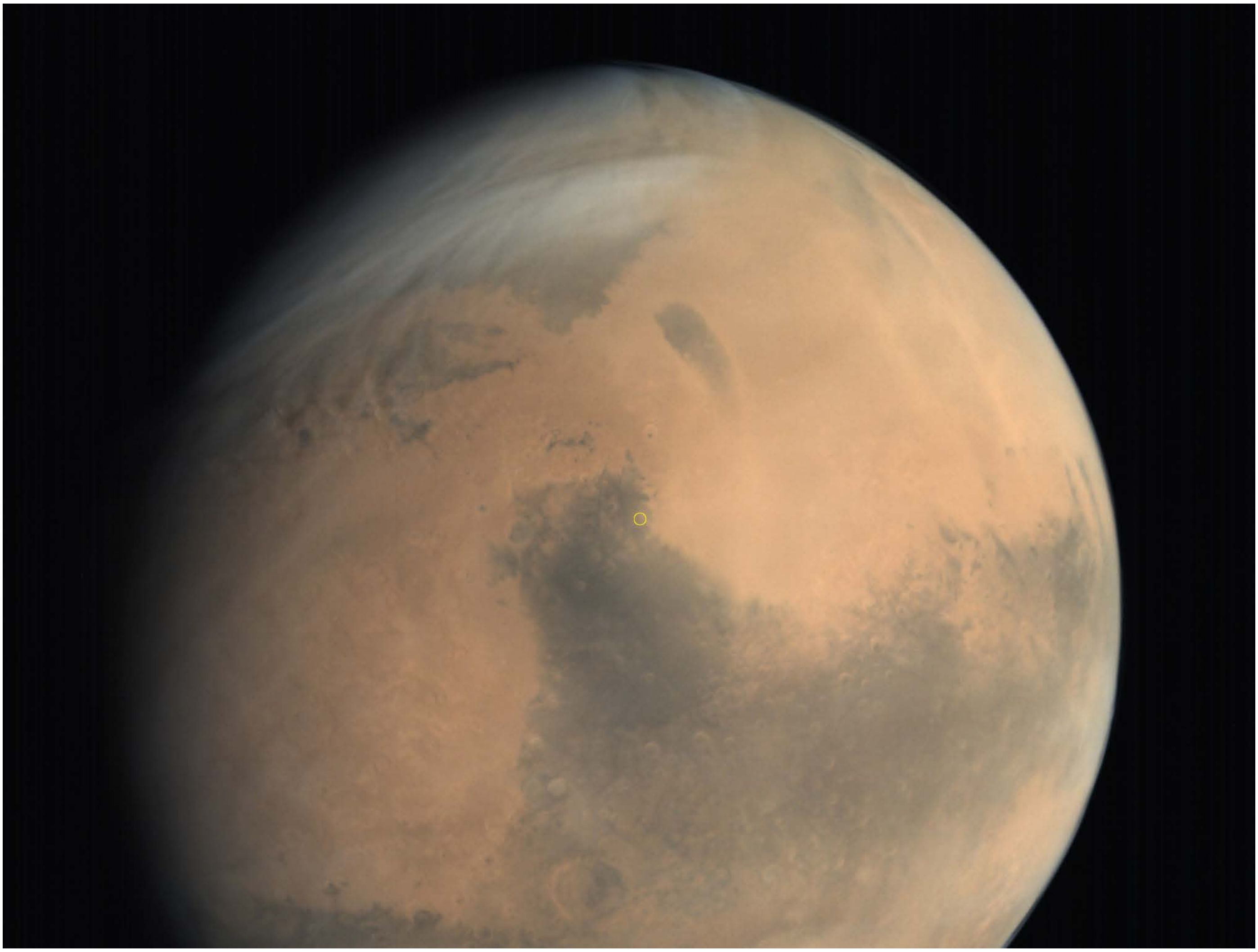
The postdoc, who asked not to be named to protect her career, told *Nature* that starting in the autumn of 2018, Pandolfi told her he was in love with her and frequently sent her personal e-mails declaring his feelings. He also organized "too many one-to-one meetings where he talked about his feelings for me", she says. She adds that she told him to keep their relationship professional, but to no avail. "It was embarrassing, horrible and I was not able to work." The postdoc was transferred to a different research group after she reported his behaviour in April 2019.

Pandolfi says the evaluation service deemed him fit to resume work and oversee his research but recommended professional and psychological coaching. He says the coaching, which he began in September, was useful.

Image questions

Pandolfi denies all allegations of misconduct in his research papers. Most allegations suggest that images of molecular assays contain duplication or inappropriate altering. On 29 May, Baumeister asked for advice from Enrico Bucci, a science-integrity expert in Samone, Italy. Bucci examined 33 papers co-authored by Pandolfi. He considers 13 of these studies to have serious problems. Pandolfi is corresponding author on eight of these.

Pandolfi says that his laboratory is extremely careful to maintain and present its data correctly. But he says he will look again at the papers under scrutiny and make any corrections that might be necessary. "I take this very seriously," he says.



EMILY LAKDAWALLA/ISRO/ISSDC

The US rover Perseverance will land in Mars's Jezero Crater (circled in yellow).

ALL ABOARD TO MARS

The United States, China and the United Arab Emirates all plan groundbreaking trips to the red planet — a notoriously dangerous destination for space missions.

By Alexandra Witze, Smriti Mallapaty and Elizabeth Gibney

Three times in the coming month or so, rockets will light their engines and set course for Mars. A trio of nations — the United States, China and the United Arab Emirates (UAE) — will be sending robotic emissaries to the red planet, hoping to start new chapters of exploration there.

Each mission is a pioneer in its own right. The United States is sending its fifth rover, NASA's most capable ever, in the hope of finding evidence of past life on Mars and collecting a set of rocks that will one day be the first samples flown back to Earth. China aims to build on its lunar-exploration successes by taking one of its rovers to Mars for the first time. And the UAE will be launching an orbiter — the first interplanetary mission by any Arab nation — as a test of its young but ambitious space agency.

It is far from a given that all these missions will make it; Mars is notorious as a graveyard for failed spacecraft. But if they do, they will substantially rewrite scientific understanding of the planet. The two rovers are heading for parts of Mars that have never been explored, and the UAE's orbiter will track the changing Martian atmosphere (see 'Mars invasion').

The teams behind the missions have managed to keep their projects on track despite the coronavirus pandemic that has derailed so many other plans, including a European–Russian Mars mission that has been delayed by two years. When the three craft lift off in the next few weeks, they will give residents of Earth a chance – however briefly – to look upwards and beyond the problems at home.

NASA's hunt for rocks

NASA hopes that its mission to Mars – a six-wheeled, three-metre-long rover named Perseverance – will be the start of a much bigger journey. If all goes to plan, Perseverance will extract and store samples of Martian rocks that a future mission will one day pick up and bring back to Earth, possibly by 2031. It would be the first-ever sample return from Mars.

That means the stakes for Perseverance are sky-high. NASA's four previous Mars rovers – 1997's Sojourner; 2004's Spirit and Opportunity; and 2012's Curiosity – were all about exploration. Mission controllers could take their time driving those machines, sidling them up to interesting-looking rocks or setting them off across vast plains. But Perseverance will arrive at Mars with the focused task of identifying and collecting a broad range of rocks representing the geological history of the area. And it is supposed to fulfil that mission in one Mars year – nearly two Earth years. Whatever the rover picks up will help to shape the course of Mars science for decades to come.

Most significantly, Perseverance represents the best chance yet for scientists to learn whether life ever arose on the red planet. If it collects the right kinds of rock, then scientists in laboratories back on Earth might be able to tease out signatures of Martian life.

"This mission gives us the first opportunity to take fundamental questions about whether there was or wasn't life on Mars to the next level," says Sherry Cady, an astrobiologist at the Pacific Northwest National Laboratory in Sequim, Washington, who is not directly involved with the mission.

Perseverance will do this with a suite of scientific instruments for poking and probing the Martian surface and atmosphere. It is a familiar-looking rover – basically a copy of the Curiosity rover that has been exploring Gale Crater for the past eight years. NASA's goal was to save money by using the same design with some tweaks, such as adding a system to store samples and upgrading the wheels. Despite its aim to cut costs, the mission's price tag has risen to US$2.7 billion, nearly $360 million over budget, because of problems developing some of the instruments.

The rover carries advanced versions of some of Curiosity's sensors, including a chemical analyser that blasts rocks with a laser to identify the atoms and molecules they are made of, and a sharp-eyed camera system that can zoom in on areas of interest to produce stereo and 3D pictures. Perseverance also sports an experiment that will try to produce oxygen from Mars's carbon dioxide-rich atmosphere, as a test of ways to support future human explorers. The rover has X-ray and ultraviolet spectrometers for analysing mineralogy in detail – and, for a bit more novelty, microphones for listening to Martian sounds, plus a squat, solar-powered helicopter.

Then there is the sampling system, which engineers designed from scratch. Perseverance carries 43 tubes in its belly. When it encounters a rock that mission scientists want to sample, the rover will reach out its 2.1-metre-long robotic arm and drill a sample about the size of a penlight: 60 millimetres long and 13 millimetres across. The sample goes into a tube and is sealed. Eventually, once Perseverance has filled at least 20 of its tubes, it will cache them on the surface of Mars until some future, yet-to-be-funded robot arrives to retrieve them. NASA currently plans to work with the European Space Agency (ESA) to launch a mission in 2026 that would return the rocks to Earth in 2031.

Perseverance will land in the 45-kilometre-wide Jezero Crater, just north of the Martian equator and in a spot that was once home to a lake and a river delta. That ancient delta offers a rich variety of geological landscapes – where Perseverance could collect many samples that might contain signs of past life, says Kennda Lynch, a planetary scientist at the Lunar and Planetary Institute in Houston, Texas, who has studied the Jezero landing site. Engineers at the Jet Propulsion Laboratory in Pasadena, California, which built the craft, have already mapped out several routes that it could take around the delta, covering on the order of 15 kilometres. Cady says the rover will do best if it first rolls around the region to survey the landscape, then returns to collect samples.

"We'll be able to cover all of Mars, through all times of day, through an entire Martian year."

Perseverance is scheduled to launch from Cape Canaveral Air Force Station in Florida between 30 July and 15 August, and land on Mars on 18 February 2021.

China's Mars debut

China has ambitious plans for its first exploration of Mars. An orbiter, lander and rover packed with 13 scientific instruments are set to launch from an island in southern China in late July. The mission, named Tianwen-1, which means 'quest for heavenly truth', will be China's deepest probe into space. When it arrives, in February next year, the mission will aim to conduct a comprehensive survey of the planet's atmosphere, internal structures and surface environment – including searching for the presence of water and signs of life.

A previous attempt by China to send an orbiter to Mars, aboard a Russian spacecraft in 2011, ended with the probes going missing. But after that loss, China has racked up a string of wins in space. In 2013, it became the third country to land a spacecraft on the Moon. And last year, a Chinese lander touched down on the Moon's far side – the first one from any country to do so. In May, China successfully test-launched a spaceship that will shuttle crew to the country's new space station – expected to be finished in 2022.

But the Mars project is in a different league from China's previous space missions, researchers at the China Academy of Space Technology in Beijing said in a 2017 paper (P. J. Ye *et al. Sci. China Technol. Sci.* **60**, 649–657; 2017). The voyage to Mars is 1,000 times longer than that to the Moon, and the planet has twice the surface gravity, an atmosphere and is littered with dense rock, which makes the effort much more risky.

The Chinese government has been tight-lipped about the mission: most public information has come from published articles and state-media reports, which omit key details about its budget, the exact launch date and where the probe will land on the planet. Scientists involved with the mission have declined interview requests until after the launch. But Wang Chi, a space physicist and director-general of the National Space Science Center (NSSC) in Beijing, said in an e-mail that the mission is moving forward as planned. "Our team is working in the Wenchang launch centre right now, and everything goes smoothly," he said, referring to the facility on Hainan Island. Wang is responsible for the scientific payloads involved in the mission, which is being led by the China National Space Administration.

If everything goes as planned, Tianwen-1 will be the first mission to successfully study the red planet with an orbiter, lander and rover. Once the combined craft reaches Mars, the hexagonal orbiter will release the lander and rover – protected by a spherical cone – into the Martian atmosphere. The Chinese team has identified two potential landing areas north of the equator on the plains of Utopia Planitia, according to a presentation made by Wei Yan of the National Astronomical Observatories in Beijing, who spoke at the European Planetary Science Congress in Geneva, Switzerland, last September.

The probe will parachute and then hover to the ground, settling on the circular lander's four legs. The rover, weighing some 200 kilograms, will then extend its solar panels, drive down a ramp and begin to autonomously explore its surroundings for the rest of its lifetime of around 90 Martian days, each of which lasts 24 hours and 37 minutes. During the rover's mission, the orbiter will act as a

MARS INVASION

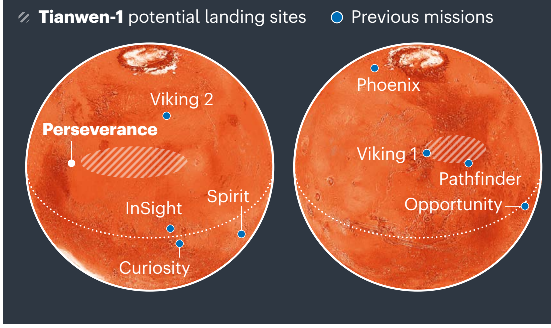
Three spacecraft heading to the red planet this year will send back an unprecedented stream of information about the alien world.

By Richard Monastersky
Design by Jasiek Krzysztofiak

Never before will such a diverse array of scientific gear have arrived at a foreign planet at the same time, and with such broad ambitions. Missions from China, the United States and the United Arab Emirates (UAE) will include two orbiters, two rovers, a stationary surface laboratory and even a helicopter. They aim to study everything from Mars's buried water deposits to the top of its atmosphere, with a particular focus on the search for life.

Landing sites

A US rover called Perseverance will land in Jezero Crater, near a delta formed by an ancient river — a prime location for finding signs of past life if it existed. China is considering several landing sites for its Tianwen-1 mission.



The escaping atmosphere

Mars once had a thick atmosphere and a significant amount of liquid water on the surface, but much of the atmosphere has leaked away over billions of years. Hope will assess how oxygen and hydrogen atoms and ions are escaping into space.

HOPE

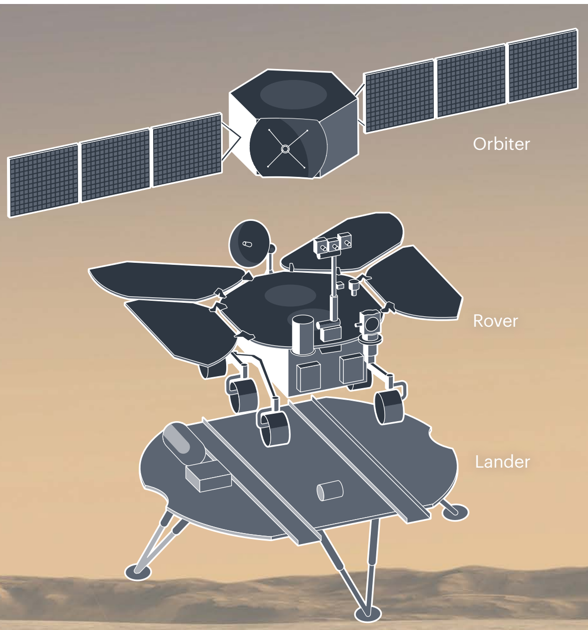
The UAE mission will travel around Mars in an elliptical orbit ranging from about 22,000 to 44,000 kilometres. It carries two spectrometers and a high-resolution imager to capture information about how the atmosphere changes over the day and throughout the seasons.

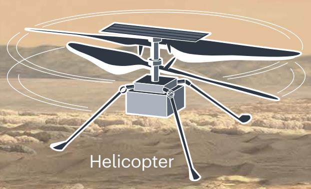
PERSEVERANCE

The US rover is a car-sized vehicle packed with seven instruments. Its main task is to collect rock samples destined to be carried back to Earth in a future mission. It will also study the planet's weather and geology, hunt for water, produce oxygen from carbon dioxide, record sounds for the first time and test a solar-powered helicopter.

TIANWEN-1

China's pioneering mission to Mars will carry an orbiter, rover and lander — it would be the first nation to achieve all three. Both the rover and orbiter have radar instruments for spotting water and ice on the surface and underground. They will also study the planet's geology and weather.

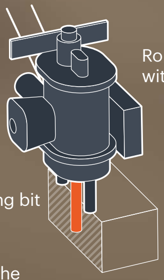




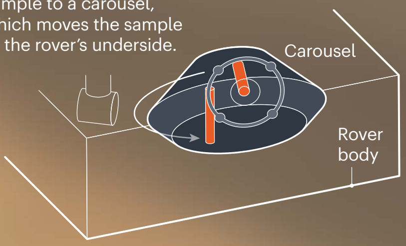
The robot geologist

Perseverance carries 43 tubes to hold rock samples collected and stored by a series of 3 robots. When the samples are eventually returned to Earth, they could provide the first definitive evidence of whether life once existed on Mars.

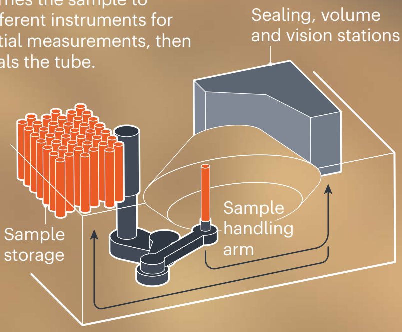
1 The 2.1-metre-long robot arm drills a thin sample of rock.



2 The arm delivers the sample to a carousel, which moves the sample to the rover's underside.



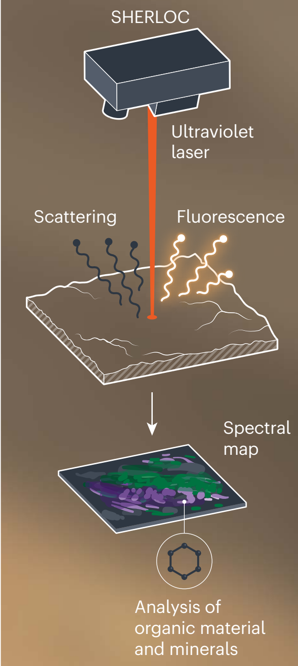
3 A second robotic arm carries the sample to different instruments for initial measurements, then seals the tube.



4 A future mission will aim to retrieve the cached samples and send them back to Earth.

Search for life

Perseverance has several instruments that will hunt for evidence of life. One of those is SHERLOC, which illuminates rocks with an ultraviolet laser and records spectra of the luminescence and reflectance. It can identify the signal of organic molecules and minerals that formed in watery environments.



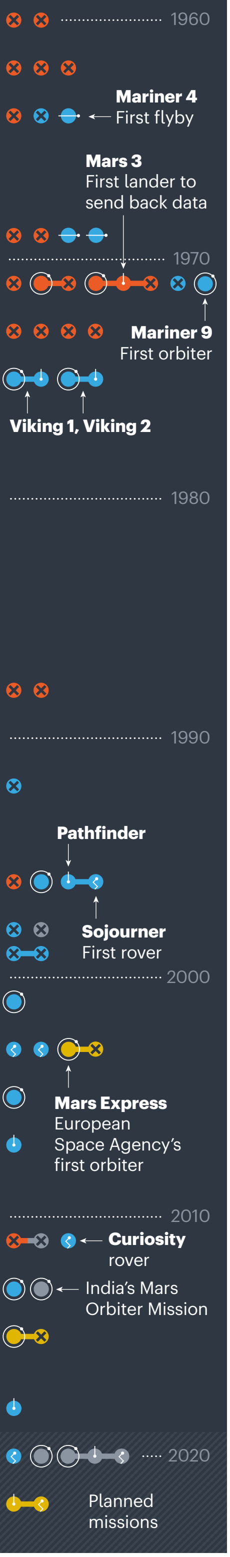
SOURCES: LANDING SITES, COMPILED BY E. LAKDAWALLA/PLANETARY SOCIETY; ATMOSPHERE GRAPHIC, ADAPTED FROM ESA/ HTTPS://GO.NATURE.COM/2ZDWJKA; HISTORICAL MISSIONS, NASA; NASA ROVER, NASA/JPL-CALTECH; MARS BASE MAP, US GEOLOGICAL SURVEY.

Historical Mars missions

● Soviet Union/Russia
● United States
● Europe
● Other

⊗ Mission failure

— Flyby
○ Orbiter
Lander
Rover

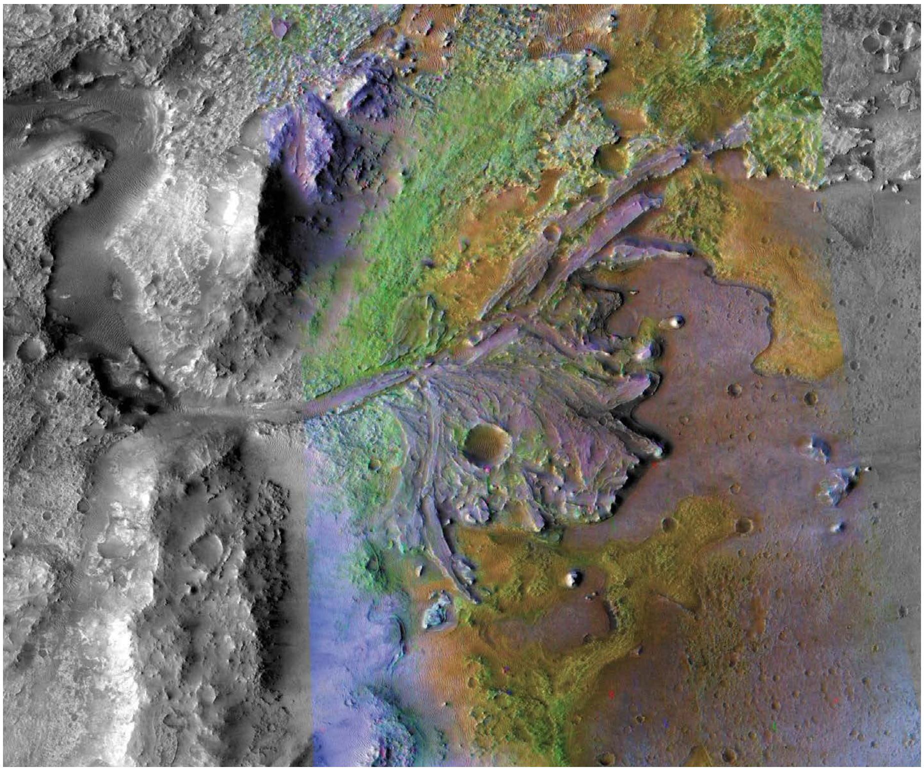


communication link, and then will move into closer orbit to survey the planet for an entire Martian year.

The Chinese team has fitted the orbiter with eight instruments, and the rover with five. The subsurface radar on the orbiter can peer 100 metres deep to map geological structures and search for water and ice. Medium- and high-resolution cameras will collect images of features such as dunes, glaciers and volcanoes, providing clues to how they formed. Both the orbiter and rover will carry spectrometers to study the composition of soil and rocks, looking especially for evidence of how water has altered geological features. The team also plans to collect atmospheric data on temperature, air pressure, wind speed and direction, as well as study the magnetic and gravitational fields on Mars.

Similar instruments have been sent on previous missions to Mars, says Raymond Arvidson, a planetary geologist at Washington University in St Louis, Missouri. But Mars is big, and has a complicated geological history, so the data generated from Tianwen-1 could inform researchers' understanding about locations not covered by existing observations, he says. "If the Chinese instruments work, produce data, and the data are shared in a manner similar to what we do, it will all be worth the effort," says Arvidson, referring to a free public archive of geoscience data collected from many previous planetary explorations that is managed by his university and NASA.

Dmitrij Titov, project scientist for ESA's Mars Express orbiter, which launched in 2003, says the Chinese orbiter could outlast some of the veterans that might be nearing the end of their life, including Mars Express, NASA's Mars Reconnaissance Orbiter and the Mars Atmosphere and Volatile Evolution orbiter, known as MAVEN. Continuous monitoring of the planet will benefit the community at a time when many other space agencies will be busy building sample-return missions, says Titov. In fact, China has its own plans to collect and bring back samples from Mars by 2030.



NASA/JPL/JHUAPL/MSSS/BROWN UNIV.

Perseverance will explore an ancient river delta where water once flowed on Mars's surface.

UAE's interplanetary hope

The United Arab Emirates had big dreams when it decided to shoot for Mars with its first probe to go beyond Earth orbit. So it chose the name 'Hope' for the orbiter, which is set to blast off from the Tanegashima Space Centre in Kagoshima, Japan, during a three-week window starting on 15 July.

If successful, the Emirates Mars Mission (EMM) will not only mark the first interplanetary venture of any Arab nation, but also produce the first global weather map of Mars. Although previous probes built up a picture of the planet's atmosphere from orbits that allowed them to monitor each part of the planet at limited times of day, Hope's huge elliptical orbit will enable the orbiter to observe big chunks of Mars under both day- and night-time conditions, covering almost the entire planet in each 55-hour orbit. "We'll be able to cover all of Mars, through all times of day, through an entire Martian year," says Sarah Al Amiri, science lead for the project and the country's minister for advanced sciences. The probe's visible-light camera and infrared spectrometer will study Martian clouds and dust storms in the lower atmosphere. Its ultraviolet spectrometer will monitor gases in the upper atmosphere. "This is the first mission that will give a global picture of the dynamics of the Mars atmosphere," says Hessa Al Matroushi, a member of the EMM science team.

During its two-year mission, Hope will track daily weather variations and changing seasons. As well as helping to prepare for future human missions, it should reveal how atmospheric conditions cause hydrogen and oxygen to escape into space. This could help scientists to understand Mars's climate and how it lost its once-thick atmosphere. The team worked with international collaborators to come up with its science goals, and the data will be made available to the international community with no embargo period, says Al Amiri. "The Emiratis were very keen to make this not just a technology demonstrator, but make it contribute to the scientific understanding of Mars," says Richard Zurek, who is the chief scientist for the Mars Program Office at NASA's Jet Propulsion Laboratory.

An interplanetary spacecraft was a significant leap in capability for the UAE, which hired seasoned engineers from previous NASA missions, mainly at the University of Colorado Boulder. The partnership has an explicit goal of transferring know-how to the team at the Mohammed Bin Rashid Space Centre, with whom the engineers worked on each element of the mission. "The reality is we are a young country and we couldn't do anything that we did without partners and international collaboration," says Ahmad Belhoul, minister for higher education and chair of the UAE Space Agency.

Unusually for an interplanetary project, the idea for the mission came not from scientists but from the government itself — and with a non-negotiable deadline of 2 December 2021, the country's 50th anniversary. Picking such an audacious task was designed not only to inspire young people in the region but also to kick-start the UAE's move to a knowledge-based economy, says Omran Sharaf, project director for the EMM.

And the mission is already having an impact, with universities offering five new undergraduate courses in pure sciences and growing enthusiasm for space among Emirati children.

In many ways, even if Hope blows up on the launch pad, the mission would be a success, says Al Amiri, who quickly reconsidered that point. "My heart just skipped a beat just thinking about it."

Alexandra Witze writes for *Nature* from Boulder, Colorado. **Smriti Mallapaty** is a senior reporter in Sydney, Australia, and **Elizabeth Gibney** is a senior reporter in London.

MANUEL VELASQUEZ/GETTY

A woman sells face masks in Mexico City.

Women are most affected by pandemics — lessons from past outbreaks

Clare Wenham, Julia Smith, Sara E. Davies, Huiyun Feng, Karen A. Grépin, Sophie Harman, Asha Herten-Crabb & Rosemary Morgan

The social and economic impacts of COVID-19 fall harder on women than on men. Governments need to gather data and target policy to keep all citizens equally safe, sheltered and secure.

Women are affected more than men by the social and economic effects of infectious-disease outbreaks. They bear the brunt of care responsibilities as schools close and family members fall ill[1,2]. They are at greater risk of domestic violence[3] and are disproportionately disadvantaged by reduced access to sexual- and reproductive-health services. Because women are more likely than men to have fewer hours of employed work and be on insecure or zero-hour contracts, they are more affected by job losses in times of economic instability[2].

There has been a "horrifying global surge in domestic violence" since the start of the COVID-19 lockdowns, said United Nations secretary-general António Guterres in early April. Malaysia, for example, reported 57% more calls to domestic-abuse helplines between 18 March and 26 March. Moreover, sexual- and reproductive-health clinics are closing worldwide. Some US states have restricted access to abortions[4].

It is all too familiar. During outbreaks of Ebola and Zika viruses in the past few years, women's socio-economic security was upended[2], and for longer than men's[1]. During the West African Ebola outbreak of 2014–16, for example, quarantines closed markets for food and other items. This destroyed the livelihoods of traders in Sierra Leone and Liberia, 85% of whom were women[5]. Men lost jobs, too, but 63% had returned to work 13 months after the first case was detected. For women, the proportion was 17% (ref. 2).

At the same time, too little is known about the differential impacts of outbreaks on men and women. And that can leave political and policy responses flying blind. Only a minority of governments collect and share basic, disaggregated sex and gender data on cases of infectious disease and the socio-economic impacts of the response to outbreaks. Analysis remains high level, often conducted after the fact and with incomplete information (go.nature.com/2a9gtja). This time, gaps must be plugged.

Here, we call for COVID-19 research, response and recovery efforts that are tailored to support women (see 'How to minimize the gendered impact of COVID-19'). The three priorities are to tackle domestic violence; ensure access to sexual- and reproductive-health services; and support women's livelihoods.

We recognize that gender is neither binary nor fixed; that the pandemic differentially affects non-binary and transgender people (go.nature.com/2zym8jc); and that gender in global health intersects with other social stratifiers such as ethnicity and race, religion, location, disability and class[6]. Therefore, beyond what we set out here, efforts to reduce the differential effects of COVID-19 must explore these intersections of marginalization and vulnerability.

Domestic violence

Domestic abuse has increased around the world since social isolation and lockdown measures for COVID-19 began[7], affecting women and girls more than men[8]. In March, the media reported that a woman was killed at the hands of a partner every 29 hours in Argentina — that's around 4 more women than the monthly average (go.nature.com/3evkopw). The official statistics are yet to be reported, and are often unreliable because reports tend to omit the victim–perpetrator relationship and motive[9]. With isolation measures restricting the movement of women and their privacy, many will be struggling to access help. Cases of domestic abuse are likely to increase as COVID-19 continues and data are collected[3].

How to minimize the gendered impact of COVID-19

Steps must be taken at three stages on domestic violence, sexual and reproductive health, and jobs.

Before. States must learn from problems and solutions during previous outbreaks, and from the first wave of COVID-19. In May, the World Health Organization issued a briefing document considering the gendered effects of COVID-19 (go.nature.com/3hubc4k). It must follow up with guidelines for best practice.

Such guidance should be integrated into domestic preparedness strategies, detailing which budget lines and indicators to track in national data sets, such as disaggregated case rates, morbidity, mortality, unemployment, crime and so on. For example, of people who have died from COVID-19 in the province of Quebec, Canada, 54% are women, where they make up the majority of care workers and care-home residents. This differs from global statistics, which show more deaths in men. In Kenya, a survey found that more women than men reported a complete loss of income or employment[19]. Can other nations adapt their policies accordingly?

During. Policymakers must accept that outbreaks affect groups differently. Governments must collect intersectional gender-disaggregated data across every aspect of the national response, from incidence and death rates, social protection and employment schemes, to accessing non-pandemic-related health services. Rapid multidisciplinary research on the gendered impact of the virus must be funded and fast-tracked into policy and strategy, and must be supported during the recovery phases. Governments must fund organizations supporting and studying those at risk from domestic violence and survivors. Sexual and reproductive health must be prioritized, protected and studied. Government policies to support livelihoods should be unconditional and broad-based, sensitive to the different impacts on men and women, and iterate as information and the situation changes.

After. Gender must be central to lessons learnt for recovery and future pandemic preparedness. Transition planning must appreciate the wider impacts on domestic abuse, livelihoods and sexual and reproductive health. For example, governments should consider how staged return-to-work policies make women or men more vulnerable to a second wave of infection, and how the rapid lifting of lockdown measures might see a surge in demand from women seeking help over domestic abuse. Any long-term recovery must consider the potential consequences of the depression on the more limited employment opportunities for women, the lower value put on their labour and their economic autonomy.

Similar patterns emerged in previous health crises. During the 2014 Ebola outbreak in Guinea, sexual and gender-based violence rose by 4.5% compared with pre-outbreak levels, according to the country's minister of social action, women and children. Last year, a study in Ebola-affected regions of the Democratic Republic of the Congo (DRC) showed that women and girls reported increases in sexual and domestic violence after the outbreak started in 2018 (go.nature.com/3duubsx).

Countries' efforts on the issue in the current pandemic vary widely. In some, it has not been addressed at all — in Kazakhstan, for example, where domestic violence is not a criminal offence[10]. And Hungary declared in May that it would not ratify the Istanbul Convention targeting violence against women, leaving women without protection from domestic abusers (go.nature.com/3ewmmpg).

By contrast, other nations braced for the onslaught. Italy increased the number of domestic-abuse helplines and set up clandestine notification protocols at pharmacies (go.nature.com/2vfxj5f). Australia boosted funding for anti-violence organizations, including those that offer safe accommodation. Kenya bolstered telephone counselling services for those facing domestic violence or the threat of it (go.nature.com/3dbvubn).

To identify where such interventions can

PATRICK BAZ/ABAAD/AFP/GETTY

Messages against domestic violence hang outside an apartment block in Lebanon.

prevent most harm, there is an urgent need to collect data using a variety of methods. This should be done during and after the outbreak, and should focus on what causes violence and where. Because domestic violence is widely under-reported, innovative methods are required.

Data gathering poses many challenges, particularly during a crisis such as COVID-19. Governments and researchers must work with survivor organizations to understand trends and impacts, changes in contexts and the socio-political dynamics. For example, how are levels of violence changing in response to lockdown or unemployment? To capture the stories of women affected by violence, whose experiences might not be apparent in official statistics, researchers will need to use qualitative methods such as interviews with community leaders, health-care providers and the women themselves[11]. Examples of best practices should be identified and shared to inform future responses to outbreaks.

Sexual and reproductive health

Global health emergencies limit and disrupt sexual- and reproductive-health services; COVID-19 is no different. This dangerous curtailment of women's rights and well-being slows progress towards achieving the UN Sustainable Development Goal on gender equality. Yet, as of 9 June, the World Health Organization's COVID-19 Strategic Preparedness and Response Plan had provided no recommendation on how resources should be channelled to provide safe abortion and ensure the supply of contraceptives.

With governments left to chart their own paths, the consequences have been grim. Contraceptives are still out of stock in Indonesia, Mozambique and many other countries. Abortions in Italy were cancelled, and are still not happening in some hospitals. Coupled with the increase in sexual violence and domestic abuse that happens in outbreaks, these problems reduce the autonomy and self-determination of women and girls, and can damage their health and well-being.

After the Ebola outbreak in Sierra Leone in 2014, some studies estimated that teenage pregnancies were 23% higher than in the previous year[12]. Restrictions on abortions do not necessarily limit demand[13]. Driven underground, these services become unsafe. During the 2016 Zika outbreak — a virus that affects fetal development, manifesting in babies with abnormally small heads, or microcephaly — no national policy changed to increase access to reproductive-health services[14]. As a consequence, women in the Zika epicentre — in Brazil, Colombia and El Salvador — told us, in work currently under peer review, that they sought unsafe abortions through providers they found online, feminist groups and the black market. Because abortion is illegal in most states where Zika was prevalent, there are no official statistics on it.

Government policies on abortion during the current pandemic differ widely, and will lead to different outcomes for women. For example, England changed its legislation in March to permit medical abortion at home through the use of pills (mifepristone and misoprostol) to terminate pregnancy after online consultation with a physician. Conversely, the states of Texas, Ohio, Iowa, Oklahoma and Alabama have further restricted access to abortion, deeming it a non-essential service[4].

The family-planning organization Marie Stopes International estimates that there could be up to 2.7 million extra unsafe abortions performed as a consequence of COVID-19.

In the short term, policymakers should take three urgent steps. First, they should make contraceptives freely available at pharmacies. Second, they should permit medical abortions at home, in consultation online with a health professional. Third, policymakers should develop a minimum initial service package for sexual and reproductive health to be implemented at the start of every humanitarian crisis. It should ensure access to contraception, obstetric and newborn care, and safe abortion care.

The package should be implemented at the start of every humanitarian crisis. The increases in sexual and domestic violence during the DRC Ebola outbreak reveals the difficulties of prioritizing sexual and reproductive health during emergencies, when health-care systems are already strained. The Inter-Agency Working Group on Reproductive Health in Crises in New York City details what governments and donor organizations should provide to women and girls to meet reproductive-health needs. For example, women are more likely to use services in locations that are less risk-prone, such as in community centres, rather than in hospitals, which are often seen as disease hotspots.

In the longer term, researchers should consider the effects of reduced access to sexual- and reproductive-health services during the pandemic. Comparing how women engage with services during a crisis and normal periods can help to analyse fertility rates or barriers to health care. For example, women changed their reproductive decisions because of the risks posed by Congenital Zika Syndrome, but this was not uniform across society. Fertility declined more in higher socio-economic groups than in low-income groups[15]. Such insights allow governments to target programmes to where they are most needed.

Livelihoods

COVID-19 is decimating livelihoods across the world. The Organisation for Economic Co-operation and Development, the African Union and the International Monetary Fund all predict potentially frightening consequences for national, regional and global economies.

By 27 March, 84 countries had adopted fiscal measures to mitigate the economic effect on households[16]. By 12 June, the number had risen to 195. Most governments increased either the coverage or payout amounts from existing social-protection schemes. Forty-seven countries have made cash-transfer programmes more flexible by waiving conditions such as the requirement for children to attend school and for women to attend ante- and postnatal appointments (such as in the Philippines). Some, such as Armenia, have provided home delivery of payments for elderly people. And 64 governments have amended unemployment benefits; 49 have adopted paid sick-leave interventions[16].

So far, only 16 countries have reported new or amended social-protection measures that make reference to women. Pakistan, for example, has increased cash transfers to women who are already receiving financial assistance from the state. Algeria has introduced paid leave for women who are pregnant, have chronic diseases or are taking care of children. Togo is providing women with US$21 per month, whereas men receive $17: President Faure Gnassingbé specified in April that this was because women are “more directly involved in nurturing the entire household”. Canada has increased its national childcare benefit, which is directed to mothers unless otherwise requested. These policies recognize the specific and increased burden that COVID-19 is having on women because of social expectations around caring responsibilities.

> **“With governments left to chart their own paths, the consequences have been grim.”**

Yet, most countries’ interventions overlook the fact that the economic consequences are likely to be worse for women.

Measures do not sufficiently cover workers in the gig or informal economy, such as street vendors or those on zero-hour contracts. They are at particular risk, because they lack the social protections of those who are formally employed. In particular, in low- and middle-income countries, 92% of women and 87% of men work in the informal economy[17]. Although the difference between these proportions is small, women tend to work in positions that leave them more open to exploitation and abuse, such as in domestic work, home-based work or by contributing to family businesses[17]. High-income countries are not immune to these trends: data from the European Institute for Gender Equality suggest that 26.5% of women employees in the European Union work in precarious employment, compared with 15.1% of men (go.nature.com/3eaabbt).

Women have higher representation in the sectors that are now laying off employees, such as hospitality, travel, education and retail (see, for example, go.nature.com/2zalzme). Many women have had to stop any casual work to meet care duties during lockdown.

Much broader measures are urgently needed for these workers and their families. By 22 May, just 94 of the 190 countries or regions for which information was available had reported commitments to support informal workers financially, leaving millions at risk[16]. Spain has committed to a universal basic income that will protect all workers. By contrast, Hong Kong gives universal payments only to permanent residents. This will not cover the 5% of the city’s population who are migrant domestic workers – mainly women[18]. Australia’s JobKeeper programme pays wage subsidies to salaried employees during the pandemic, but not to casual workers – who are more often women (go.nature.com/2zalzme).

In the short term, governments should focus on help for informal and casual workers. For example, removing requirements that a person must have had previous taxable income to benefit from COVID-19-related relief, and ensuring that unemployment benefits and statutory sick pay meet basic needs.

This is also a time for innovation. New Zealand, for example, is suggesting a four-day working week to mitigate rising unemployment, to support a better work–life balance and to boost local tourism. The idea comes from a well-being budget that it introduced last year (go.nature.com/2bjt1qa).

To inform the long road out of this global depression, we need to monitor the real-world impact of policies on the hardest hit in real time, so that strategies can be adjusted if necessary. Such research requires sex-disaggregated data on the workforce. The UK government, for example, suspended collection of data on the gender pay gap during the pandemic because it was deemed non-essential. Such information is more crucial now than ever.

Context is key

Broad-brush comparisons of vulnerabilities to COVID-19 responses are to be treated with caution. Gender and its impacts are context-specific, and vary between and within countries. The data collected in other health emergencies in Liberia, Yemen or Brazil can suggest trends. But data sets are often incomplete, and the nuances are highly dependent on race, religion, ethnicity, location, disability and class[6]. Addressing some of the issues that women face in outbreaks highlights a broader landscape of inequalities. Policymakers must consider and support all those at the margins.

Our critics might advocate for other priorities. We’re calling on governments to use evidence to ensure that all their citizens have an equal chance of safety, shelter and security. And when the pandemic ends, addressing gender inequality must be at the heart of the broader programme to ‘build back better’.

The authors

Clare Wenham is an assistant professor of global health policy at the London School of Economics and Political Science, UK.

Julia Smith is a research associate in the Faculty of Health Sciences at Simon Fraser University, Burnaby, Canada. **Sara E. Davies** is a professor of international relations at the School of Government and International Relations, Griffith University, Brisbane, Australia. **Huiyun Feng** is a senior lecturer at the School of Government and International Relations, Griffith University, Brisbane, Australia. **Karen A. Grépin** is an associate professor at the School of Public Health, University of Hong Kong, Hong Kong. **Sophie Harman** is a professor of international politics in the School of Politics and International Relations, Queen Mary University of London, UK. **Asha Herten-Crabb** is a PhD student in the Department of International Relations at the London School of Economics and Political Science, UK. **Rosemary Morgan** is an assistant scientist in the Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
e-mail: c.wenham@lse.ac.uk

1. Human Rights Watch. *Neglected and Unprotected: The Impact of the Zika Outbreak on Women and Girls in Northeastern Brazil* (Human Rights Watch, 2017).
2. Bandiera, O. *et al. The Economic Lives of Young Women in the Time of Ebola: Lessons from an Empowerment Programme (Working Paper F-39301-SLE-2)* (International Growth Centre, 2018).
3. Peterman, A. *et al. Pandemics and Violence Against Women and Children: Working Paper 528* (Center for Global Development, 2020).
4. Bayefsky, M. J., Bartz, D. & Watson, K. L. *N. Engl. J. Med.* **382**, e47 (2020).
5. Kabia, A. B. 'Women in Sierra Leone's Labour Market.' *China Daily* (9 February 2016); available at https://go.nature.com/3fswaiz
6. Davies, S. E., Harman, S., Manjoo, R., Tanyag, M. & Wenham, C. *Lancet* **393**, 601–603 (2019).
7. Roesch, E., Amin, A., Gupta, J. & García-Moreno, C. *Br. Med. J.* **369**, m1712 (2020).
8. UK Office for National Statistics. *Domestic Abuse Victim Characteristics, England and Wales: Year Ending March 2019* (ONS, 2019).
9. World Health Organization. *Understanding and Addressing Violence against Women: Femicide* (WHO, 2020)
10. Klugman, J. *Gender Based Violence and the Law* (World Bank, 2017).
11. Fraser, E. *Impact of COVID-19 Pandemic on Violence against Women and Girls.* VAWG Helpdesk Research Report No. 284 (UK Department for International Development, 2020).
12. United Nations Population Fund. *Recovering from the Ebola Virus Disease: Rapid Assessment of Pregnant Adolescent Girls in Sierra Leone* (UNFPA, 2018).
13. Sedgh, G. *et al. Lancet* **388**, 258–267 (2016).
14. Wenham, C. *et al. Glob. Health* **15**, 49 (2019).
15. Marteleto, L. J., Guedes, G., Coutinho, R. Z. & Weitzman, A. *Demography* https://doi.org/10.1007/s13524-020-00871-x (2020).
16. Gentilini, U. *Social Protection and Jobs Responses to COVID-19: A Real-Time Review of Country Measures* (World Bank, 2020).
17. Bonnet, F., Vanek, J. & Chen, M. A. *Women and Men in the Informal Economy: A Statistical Brief* (WIEGO/ILO, 2019).
18. Man, J. 'In Hong Kong, a setback for domestic-worker rights.' *Time* (29 March 2012).
19. Abuya, T. *et al.* 'COVID-19-related knowledge, attitudes, and practices in urban slums in Nairobi, Kenya.' Version 13.0 Harvard Dataverse https://doi.org/10.7910/DVN/VO7SUO (2020).

Sustainable Development Goals: pandemic reset

Robin Naidoo & Brendan Fisher

COVID-19 is exposing the fragility of the goals adopted by the United Nations — two-thirds are now unlikely to be met.

As COVID-19 batters the world and its economy, it's time to rethink sustainable pathways for our planet. Rosy hopes that globalization and economic growth would bankroll waves of green investment and development are no longer realistic. It's unlikely there will be enough money or attention to banish poverty and inequality, expand health care and overturn biodiversity loss and climate change, all by 2030.

The SARS-CoV-2 virus has already killed more than 512,000 people, disrupted the livelihoods of billions and cost trillions of dollars. A global depression looms. The United States and other nations are gripped by protests against structural inequality and racism. And geopolitical tensions between superpowers and nuclear states are at levels not seen for decades.

Things were different back in 2015, when the United Nations adopted 17 Sustainable Development Goals (SDGs) to improve people's lives and the natural world by 2030. It was arguably one of humanity's finest moments — the whole planet signed up. Many national budgets were flush with funds. Governments agreed ambitious treaties, including the Paris climate agreement, the Sendai framework on disaster risk reduction and the Addis Ababa plan for financing development.

Five years on, as the UN celebrates its 75th anniversary, that mood of optimism has gone. In other words, the very foundations on which the SDGs were built have shifted.

The success of the SDGs depends on two big assumptions: sustained economic growth and globalization. COVID-19 has torn these to shreds. The global economy is expected to contract by at least 5% this year, and the timeframe for its recovery is years, not months, if the past is any guide. Industrialized countries struggling to support their own citizens will not bankroll the development of others. Overseas development aid could drop by US$25 billion in 2021. The United States has announced its withdrawal from the World Health Organization. Increasing the scale of human activity on the planet looks foolish when it could open wells of new diseases once hidden in the wild, similar to COVID-19.

Governments have basic worries. Food security is under threat, because farm workers are unable to travel to harvest crops; prices of rice, maize (corn) and wheat are rising. The UN World Food Programme has just doubled its estimate of the number of people who are likely to face acute food shortages this year, to 265 million. Demand for cash crops, such as Kenya's flower exports, has stalled. Ecotourism has collapsed. Even oil-rich developing countries such as Nigeria, Africa's most populous nation, cannot sell their resources profitably in the global slowdown.

And the world will face further stressors in the next decade. More pandemics, yes, but also extinctions and the continued degradation of the ecosystems on which all life depends. Storms, wildfires, droughts and floods will become more frequent owing to climate change. Geopolitical unrest might follow. Mounting costs to address these will divert yet more funding from existing SDG targets. Last year alone, the United States experienced 14 separate billion-dollar disasters related to climate change.

COVID-19 is demonstrating that the SDGs as currently conceived are not resilient to such global stressors. As the UN's High-level Political Forum on Sustainable Development meets (virtually) this week, delegates must chart a new course for the SDGs. As the world recovers from this pandemic, the forum must establish a few clear priorities, not a forest of targets. It should also consider which goals can be achieved in a less-connected world with a sluggish global economy.

Slow or worse

Progress across the SDGs was slow even before COVID-19. Now, it's even more likely that many of the 169 targets will not be met by 2030. Worse, some could even be counterproductive (see 'COVID-19 impacts on Sustainable Development Goals'). Two-thirds of the 169 targets are either under threat as a result of this pandemic or not well-placed to mitigate its impacts (see Supplementary information).

NARINDER NANU/AFP/GETTY

Migrant labourers in Uttar Pradesh protest against the lack of food in a slum area after the government eased lockdown measures.

Some might even amplify problems. Ten per cent of the SDG targets could worsen the impacts of future pandemics.

The goal of good health is the most obvious casualty. Clinics everywhere are stretched by COVID-19, and resources are in short supply. The impacts are spreading to all areas of health care. For example, the UN children's charity UNICEF warns that up to 116 million newborns and mothers will experience inadequate services in the coming months.

With travel off the table, tourism is suffering. The Organisation for Economic Co-operation and Development (OECD) estimates that the number of international tourists will drop by 60% this year, reducing tourism's contribution to global gross domestic product (GDP) and affecting countries where it is a substantial part of the national economy. These include Namibia, where tourism centred around charismatic species such as lions, leopards and elephants contributes 10% of the country's GDP. Flowing more than $11 million annually into local communities has also helped to dramatically increase wildlife populations, including those of elephants that are declining elsewhere in Africa. But without tourism, poaching has reappeared – in April, two endangered black rhinoceroses (*Diceros bicornis*) were killed on community lands for the first time in almost three years.

Goals and targets that rely on a growing global economy will not be met. For example, making energy affordable and clean will require the creation of new markets and financing. Boosting industry, innovation and infrastructure will require extra investment. Even before COVID-19, financing for the SDGs was $2.5 trillion short.

> **"Cash-strapped governments need to focus on a few broad strategic goals."**

Concerns over targets that conflict with one another have been raised before[1]; now, they are pressing. For example, improving the transport network in developing countries is a key focus of the goal for industry, innovation and infrastructure. Yet, extending roads into wilderness puts more people in the path of new pathogens. Construction is damaging forests and other fragile environments across the tropics[2], counter to the goal to protect life on land. A proposed highway through Africa's Serengeti region would cause irreparable harm to a protected area that generates more than $80 million per year from tourism[3]. Similarly, marine traffic pollutes the air and ocean, and puts the goal to protect 'life under water' under threat. Air travel moves people, money and ideas, but helped to spread SARS-CoV-2 rapidly around the world[4].

What to do?

Prioritize win–wins. Cash-strapped governments need to focus on a few broad strategic goals[5]. This will inevitably upset groups that support goals and targets that are de-emphasized. But sometimes that cannot be avoided.

Priorities will be hard to identify among so many diverse targets. Some have clear and quantitative aims – for example, to eradicate extreme poverty for all people everywhere by 2030, such that no one is living on less than $1.25 a day. Others are more diffuse – such as promoting public procurement practices that are sustainable, in accordance with national policies and priorities.

Fortunately, some of the goals underpin or interact positively with many others[1,6].

RUNGROJ YONGRIT/EPA/SHUTTERSTOCK

Illegal trade in endangered species, such as pangolins, increases the risk that people are exposed to viruses from the wild.

And some become more urgent in light of COVID-19. For example, 30 (18%) of the targets would help to lessen the likelihood of another global pandemic. Reducing wildlife trafficking and the supply and demand of illegal wildlife products, for instance, would reduce the probability that new viruses would transfer to humans[7]. Three further targets – achieving universal health coverage, bolstering the health workforce and strengthening the capacity of early-warning systems for global health risks – will slow the cascading impacts of COVID-19 in low-income nations.

Experts in decision science, cost-benefit analysis and socio-ecological systems should complement political representatives in determining which SDG targets should be prioritized. The Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) shows how information from various scientific disciplines can support difficult decisions on climate-change mitigation and adaptation options[8].

Decouple development and growth. COVID-19 is a stress test of our globalized economy and of our global goals for a more sustainable planet. Just as when global banking faced, and failed, a similar test in 2008, it must be a learning experience. The financing, oversight and implementation of the SDGs have to be reformed.

Sustained per-capita economic growth for all countries is itself an SDG target, which points to just how deeply its pursuit is ingrained in the modern world. But the most common measure used, GDP, is distorting – it assigns value to undesirable factors, such as dangerous jobs, traffic jams and pollution[9]. And growth cannot continue forever on a finite planet that's already over-exploited[10].

As a result, many have long argued that economies should focus on development (improving well-being) rather than on growth (increasing economic throughput)[11]. Many investors now recognize that maximizing short-term growth cannot come at the expense of clean air and water, a stable climate, peaceful communities and resilient ecosystems. Measures that bake explicit social and environmental goals into financial instruments, such as green bonds, sustainability bonds and impact investing, are growing in popularity. And this year, BlackRock of New York City, the world's leading asset-management firm, joined the Climate Action 100+ investor initiative to push the world's largest corporate greenhouse-gas emitters to take action on reducing emissions.

> "If the world's economic pie cannot increase, it must be sliced in different ways."

Other key tenets of globalization, such as the value of interconnectedness for efficiency's sake, must be questioned, too. Yes, trade, travel and telecommunications enhance the well-being of billions of people and have marshalled flows of protective equipment and technical expertise to fight the COVID-19 pandemic. But interconnections also increase the likelihood that future global pandemics will emerge and spread, and of financial contagion and the erosion of social protections for workers. Full social and environmental costs, as well as benefits, now need to be re-examined.

Slowing another form of growth – population – should also be a priority. Historical missteps mean that such discussions can be contentious. But if the world's population rises, as predicted, to 9.7 billion by 2050, it will exacerbate all other threats to sustainability. The SDG of empowering women and educating girls is thus crucial[12]. Even more important will be stabilizing population size in high-income countries, where consumption and environmental impacts are much higher than in low-income nations, and where up to 40% of pregnancies are unintended. There is therefore both a need and an opportunity to reinvigorate research and action on the scale of human activity on our finite planet.

Overhaul funding. If the world's economic pie cannot increase, it must be sliced in different ways. A short-term solution is for OECD economies to reduce perverse underwriting of enterprises that are anathema to the SDGs. For example, government subsidies to the fossil-fuel industry reached $4.7 trillion (6.3% of global GDP) in 2015. Continued reliance on fossil fuels limits success in numerous SDGs, from sustainable energy and cities to climate change and biodiversity conservation.

Another way to re-divide the pie is to rein in corporate profits. In 2019, Fortune 500 companies collectively posted profits of $1.1 trillion. This figure is 200 times the annual budget of the World Health Organization. Although many companies will face losses and bankruptcy as a result of the pandemic, there are ways to recoup funds to support the SDGs. Curbing tax avoidance is one — each year, low-income countries lose 1.3% of their GDP through tactics to avoid corporate taxes. Companies can also have explicit aims to serve the public good and the SDGs. For example, 'Certified B Corporations' balance profit with purpose. Their numbers are increasing and include, for example, food giant Danone North America and global lifestyle brand Patagonia. Bancolombia, South America's third-largest bank, and consumer goods firm Unilever invest in companies that deliver social and environmental profits.

The mind-boggling sums invested in military defence are also at odds with the global-scale cooperation to which nations committed under the SDGs. Climate change is increasingly recognized as a national-security threat, and in the wake of COVID-19, calls to treat future pandemics as such will also increase. Diverting funds from armaments to address such security threats would provide a funding pathway to the SDGs that should be encouraged. Protests in the United States are now catalysing calls to downscale and demilitarize municipal law-enforcement agencies, re-routing funds to initiatives on mental health and other social services. A similar process at the global scale would unlock immense resources.

Road ahead

UN conventions have been another casualty of COVID-19. The 2020 UN Biodiversity Conference and the 26th United Nations Climate Change conference (COP26) have been postponed until next year. Although global crises cannot afford delays, these postponements provide an opportunity for the lessons learnt from COVID-19 to be codified into the agreements that will define life on Earth into the next century.

We therefore urge the UN's High-level Political Forum to work out how and when to update the SDGs. Every goal and target should be screened according to three points: is this a priority, post-COVID-19; is it about development not growth; and is the pathway to it resilient to global disruptions?

It is our hope that 75 years from now, the tragedy of 2020 will also be remembered as a positive watershed, after which we built back better.

COVID-19 IMPACTS ON SUSTAINABLE DEVELOPMENT GOALS

SDG	Status	Example of target(s) affected
Goal 1: No poverty	Threatened* and mitigates†	Target 1.2: halve proportion of people living in poverty by 2030 Target 1.4: provide equal access to basic services
Goal 2: Zero hunger	Threatened	Target 2.3: double agricultural productivity and incomes of small-scale food producers
Goal 3: Good health and well-being	Threatened and mitigates	Target 3.8: achieve universal health coverage
Goal 4: Quality education	Threatened	Target 4.1: provide free, equitable and quality education for all children
Goal 5: Gender equality	Partially threatened‡	Target 5.4: value unpaid care and domestic work by providing public services and policies
Goal 6: Clean water and sanitation	Threatened	Target 6.1: give access to safe and affordable drinking water for all
Goal 7: Affordable and clean energy	Threatened	Target 7.3: double global rate of improvement in energy efficiency
Goal 8: Decent work and economic growth	Threatened	Target 8.1: sustain per capita economic growth
Goal 9: Industry, innovation and infrastructure	Threatened and aggravates§	Target 9.4: upgrade infrastructure and retrofit industries to make them sustainable
Goal 10: Reduced inequalities	Threatened	Target 10.1: sustain above-average income growth of the bottom 40% of the population
Goal 11: Sustainable cities and communities	Threatened	Target 11.2: give access to safe, affordable and sustainable transport systems for all
Goal 12: Responsible consumption and production	Partially threatened	Target 12.5: reduce waste generation through prevention, reduction, recycling and reuse
Goal 13: Climate action	Threatened	Target 13.A: mobilize US$100 billion annually by 2020 for the Green Climate Fund to address the needs of developing countries
Goal 14: Life below water	Partially threatened	Target 14.1: by 2025, prevent marine pollution of all kinds
Goal 15: Life on land	Threatened and mitigates	Target 15.7: end poaching and trafficking of protected species and address demand and supply of illegal wildlife products
Goal 16: Peace, justice and strong institutions	Partially threatened	Target 16.1: reduce all forms of violence and related deaths everywhere
Goal 17: Partnerships for the goals	Partially threatened	Target 17.2: developed countries should commit at least 0.7% of gross national income in overseas aid for developing and 0.15% to least-developed nations

*Most targets unachievable. †Achieving some targets would have helped prevent pandemic impacts. ‡Some targets affected. §Achieving target would have made pandemic impacts worse.

The authors

Robin Naidoo is a lead scientist with WWF-US in Washington DC, USA, and an adjunct professor at the Institute for Resources, Environment and Sustainability, University of British Columbia, Vancouver, Canada. **Brendan Fisher** is a professor in the Environmental Program — Rubenstein School of Environmental and Natural Resources and Gund Institute of Environment, University of Vermont, Burlington, Vermont, USA.
e-mails: robin.naidoo@WWFUS.ORG; brendan.fisher@uvm.edu

Supplementary information accompanies this comment at go.nature.com/3gd6ohj

1. Nilsson, M., Griggs, D. & Visbeck, M. *Nature* **534**, 320–322 (2016).
2. Laurance, W. F. & Arrea, I. B. *Science* **358**, 442–444 (2017).
3. Hopcraft, J. G. C. *et al. Conserv. Biol.* **29**, 932–936 (2015).
4. Chinazzi, M. *et al. Science* **368**, 395–400 (2020).
5. Shepherd, K. *et al. Nature* **523**, 152–154 (2015).
6. Fisher, B. *et al. Lancet Planet. Health* **3**, e112–e113 (2019).
7. Johnson, C. K. *et al. Proc. R. Soc. B* **287**, 20192736 (2020).
8. Jones, R. N. *et al.* in *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* Field, C. B. *et al.* (eds) 195–228 (Cambridge Univ. Press, 2014).
9. Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. *Mis-measuring our Lives: Why GDP Doesn't Add Up* (New Press, 2010).
10. Steffen, W. *et al. Science* **347**, 1259885 (2015).
11. Daly, H. E. *Sci. Am.* **293**, 100–107 (2005).
12. Crist, E., Mora, C. & Engelman, R. *Science* **356**, 260–264 (2017).

Readers respond

Correspondence

US visa changes leave postdocs in limbo

Six of us are postdocs from abroad who work in the United States. We are therefore deeply concerned about the latest uncertainties over US visas, announced last month (see *Nature* http://doi.org/d2h4; 2020). Coming on top of the havoc wreaked by the COVID-19 pandemic (see, for example, *Nature* **582**, 449–450; 2020), the changes mean that researchers' careers are now under grave threat.

Things were already difficult under the preceding visa regime. Those of us in this position were wary of visiting our home countries, in case our return to the United States was blocked. The US government's imposition of travel restrictions has made matters worse, thwarting even urgent trips back to our families.

Renewing a non-immigrant visa is next to impossible at the moment, and legal advice is hard to obtain because of the fluidity and confusion of the situation. Applying for a new visa is not an option, so long as embassies and consulates are closed and travel restrictions apply. Meanwhile, our projects and job applications are on hold.

We value the opportunity to work in world-class laboratories. In return, we contribute a highly motivated and affordable talent pool to our host institutions. We urge the United States to safeguard international postdocs to reinforce these mutual benefits.

Amir H. Behbahani*, California Institute of Technology, Pasadena, California, USA.
amirhb@caltech.edu
*On behalf of 7 correspondents, see go.nature.com/3f2pnyl

Landmark 100 years of climate modelling

This year marks the centenary of the seminal work on climate modelling by the Serbian mathematician and geophysicist Milutin Milanković.

In 1920, Milanković published his book *Mathematical Theory of Thermal Phenomena Caused by Solar Radiation.* This linked long-term climatic changes to astronomical factors that affect the amount of energy Earth's surface receives from the Sun. In collaboration with geophysicist Alfred Wegener and meteorologist Wladimir Köppen, Milanković used this model to determine Earth's past climatic cycles over thousands of years, culminating in the monumental and influential 1941 work *Canon of Insolation and the Ice-Age Problem*. This led to the recognition of regular changes in key astronomical parameters: the eccentricity of Earth's orbit around the Sun, and the obliquity and precession of Earth's rotational axis.

As a result, science could at last explain the distribution of ancient large-scale glacial deposits in areas that now have a temperate or warm climate. These periodic climatic oscillations, known as Milanković cycles, also explained the repetitive glaciations that occurred during Earth's history. Crucially, these cycles enabled prediction of future climate changes.

Marco Romano Sapienza Università di Roma, Rome, Italy.
marco.romano@uniroma1.it

Bruce Rubidge University of the Witwatersrand, Johannesburg, South Africa.

Deconstruct racism in medicine

COVID-19 is four times more likely to severely affect African Americans than their white counterparts (see go.nature.com/37fffny). Structural racism in our society undoubtedly contributes to this stark difference. As physician-scientists, we have a duty to break this cycle of disadvantage through our clinical work, scientific inquiry and education efforts.

Health inequity is perpetuated by social, economic and environmental disparities in the African American community. Researchers focusing on patient-oriented studies should ensure that cohorts are representative of racial demographics. Too few people of colour currently enrol in clinical trials. This stems, in part, from mistrust, after the US Public Health Service's scandalous 1932–72 Tuskegee study of untreated syphilis in Black males (see S. M. Reverby *Nature* **567**, 462; 2019). Clinical researchers now have an obligation to patients and their families to advocate and educate on the risks and benefits of participation in clinical trials.

As educators, we must also remove the unconscious bias that affects student selection and commit to mentoring students of colour. This will expand the pipeline of under-represented scientists and better equip us to tackle racial disparities in a clinical setting.

Talia H. Swartz Icahn School of Medicine at Mount Sinai, New York, New York, USA.
talia.swartz@mssm.edu

Boghuma Titanji Emory University School of Medicine, Atlanta, Georgia, USA.

Dreadlocks and discrimination

As an Afro-Colombian soil ecologist with dreadlocks, I have encountered prejudice and scepticism about my profession countless times — from airport and immigration authorities, the public at outreach events and even colleagues at conferences.

Such experiences reinforce my conviction that, as scientists, it is our professional, civic and moral duty to strongly denounce and combat systemic and structural racism and its intersectionality with other forms of oppression and discrimination. We must study its causes (see E. Culotta *Science* **336**, 825–827; 2012), as well as its tragic consequences (see, for example, A. Mesic *et al. J. Natl Med. Assoc.* **110**, 106–116; 2018).

There is also plenty of work to do to eliminate racial under-representation in science. For example, a study published earlier this year shows that PhD students from under-represented groups in the United States innovate at higher rates than do those in the majority, but that their contributions are less likely to lead to academic positions (B. Hofstra *et al. Proc. Natl Acad. Sci. USA* **117**, 9284–9291; 2020).

César Marín University of O'Higgins, San Fernando, Chile.
cesar.marin@uoh.cl

HOW TO SUBMIT

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at go.nature.com/cmchno.

Expert insight into current research

News & views

Coronaviruses

Going back in time for an antibody to fight COVID-19

Gary R. Whittaker & Susan Daniel

Efforts are intensifying to try to harness antibodies as a therapy for COVID-19. A study reveals the insights that can be gained from antibodies made by a person who had a coronavirus infection that caused the disease SARS. **See p.290**

The COVID-19 pandemic is the biggest public-health crisis in a century, and the development of medical interventions to combat the SARS-CoV-2 coronavirus is a top priority. On page 290, Pinto *et al.*[1] provide evidence needed to take one of the crucial first steps for such efforts in the developing arena of antibody immunotherapy.

The level of protection provided by the immune system in response to SARS-CoV-2 exposure and infection is a hotly debated topic[2]. It is thought that one major arm of the immune response to such infection is the development of antibodies that recognize the virus. Of particular interest are antibodies that bind to a protein on the viral surface known as the spike protein. Coronaviruses derive their name from their distinctive, crown-like (coronal) viral silhouettes, which are due to these proteins.

Antibodies that recognize and bind to the viral 'spike' can block its ability to bind the ACE2 receptor protein on human cells. An interaction between the spike protein and ACE2 is part of a process that can enable coronaviruses to enter human cells. Thus, antibodies that could hinder spike-protein function would block infection; such antibodies are termed neutralizing antibodies.

Much remains to be learnt about the immunological responses to SARS-CoV-2. Nevertheless, it is becoming clear that antibodies taken from the blood serum of people who have recovered from COVID-19 can be used for treatment by being transfused into other people who have the disease[3]. Such 'convalescent sera' approaches are highly attractive, particularly as an immediate treatment option. That's because more-conventional therapeutics, such as drugs or vaccines, are unlikely to be available for some time. A more high-tech approach to using convalescent sera is the manipulation of antibody-producing B cells taken from the blood of people who had COVID-19 or other coronavirus infections. Each B cell makes one unique antibody, and clonal populations of a B cell of interest can be used to generate an identical pool of a particular desired antibody known as a monoclonal antibody.

To accelerate the process of therapeutic development, Pinto and colleagues 'went back in time', and turned to samples of B cells collected from a person who had been infected by the coronavirus SARS-CoV. This virus, which is similar to SARS-CoV-2, caused an outbreak in 2003 of a disease called severe acute respiratory syndrome (SARS). The hope with such an approach is that the resemblance between the two viruses might mean that some antibodies that recognize SARS-CoV also recognize and neutralize SARS-CoV-2.

The 'head', or receptor-binding domain (termed S1), of the spike protein is the most accessible region of the protein for antibodies to bind to. However, this domain exists in different dynamic states, and debate has arisen over whether it is 'masked' from the immune system by a shell of carbohydrate molecules[4]. The identification of a functional antibody that targets this region is therefore not a trivial process. Pinto *et al.* combined blood cells taken in 2004 and 2013 from a person who had recovered from SARS, and searched for antibodies that could recognize SARS-CoV-2 (Fig. 1). Of the 25 different monoclonal antibodies that the authors studied, 4 recognized the receptor-binding domains of both SARS-CoV and SARS-CoV-2 spike proteins. One antibody, termed S309, was selected for further study on the basis of its high-affinity binding to this domain when tested *in vitro*.

Pinto and colleagues used cryo-electron microscopy to visualize the interaction between the S309 antibody and the SARS-CoV-2 spike protein. This revealed that S309 binds to an accessible site in the

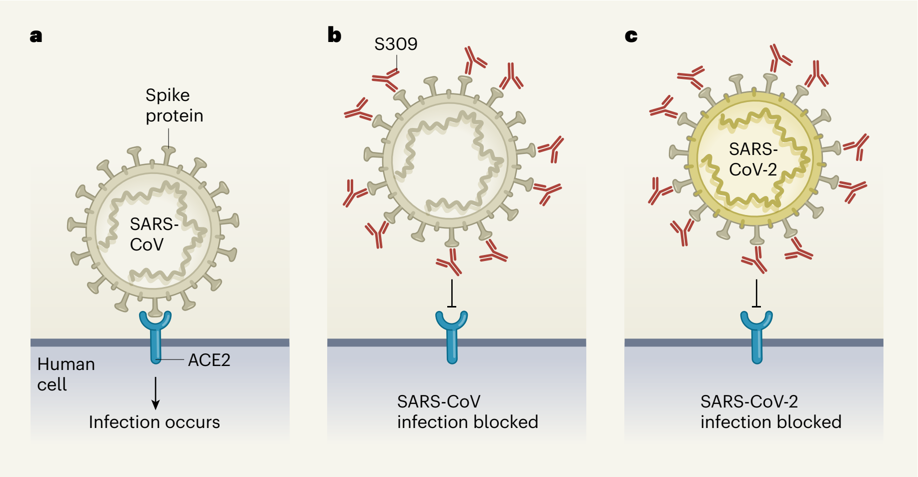


Figure 1 | An antibody that blocks coronavirus infections. Pinto *et al.*[1] have identified a human antibody that blocks infection by SARS-CoV-2, the coronavirus that causes COVID-19. The authors made this discovery by examining antibodies made by a person who had recovered in 2003 from infection with the related coronavirus SARS-CoV, which causes severe acute respiratory syndrome (SARS). **a**, Coronaviruses such as SARS-CoV infect human cells by binding to the protein ACE2. **b**, Pinto and colleagues analysed blood samples taken in 2004 and 2011 from a person who recovered from SARS, and examined antibodies made by the immune cells from the samples. They identified an antibody (named S309) that bound to the spike protein of SARS-CoV and prevented infection by this virus. **c**, The authors found that this antibody bound to a similar region of the spike protein of SARS-CoV-2 and prevented infection by the virus.

receptor-binding domain of the spike protein that has an attached carbohydrate molecule. This region is not part of the key area that directly binds to ACE2. The site that S309 recognizes is evolutionarily conserved in spike proteins across a range of bat coronaviruses (in the genus *Betacoronavirus* lineage B; subgenus *Sarbecovirus*) that have similarities to the SARS-like coronaviruses. This raises the possibility that such an antibody could have wide applicability in tackling related viruses. Not only, then, is this antibody of interest when investigating ways to manage the COVID-19 pandemic in the years ahead, but it might also be considered for use in preventing future outbreaks of related animal viruses, if they make the leap to causing infection in humans.

Ultimately, it seems unlikely that a robust treatment for COVID-19 will rely on a single antibody. Rather, as was the case for SARS, a synergistic approach combining different monoclonal antibodies in an antibody cocktail might be more effective[5]. For such approaches to move forwards, evidence of effective antibody neutralization from *in vitro* studies will be needed, along with *in vivo* data assessing how well an antibody can boost other aspects of the immune response — by enlisting other immune cells to tackle the infection, for example. There are many promising avenues to explore in these efforts.

Pinto and colleagues got a head start with their work by exploring pre-existing antibodies, and they should now have more B-cell populations to mine. Many other teams, to give just some examples[2,6–13], have also presented useful discoveries in the hunt for antibodies that can target SARS-CoV-2. The next steps will be to test individual antibodies and antibody cocktails in animal models, to determine whether they offer protection, and then to assess their safety and effectiveness in human clinical trials. An accelerated path might narrow the time lag between antibody discovery and proof-of-concept trials in humans to as little as five or six months[14].

The most recent prominent example of immunotherapy for infectious disease relates to battling the Ebola virus. In concert with vaccines and conventional, small-molecule-drug trials, the development of monoclonal-antibody therapies for Ebola has progressed rapidly. Cocktails of antibodies, beginning with one called ZMapp, that target a key Ebola viral protein called GP in two crucial regions of the protein, are continuing to be developed[15–17]. This progress in efforts to tackle Ebola gives hope for similar immunotherapy achievements in targeting SARS-CoV-2. Pinto and colleagues' work marks a major step towards that much-anticipated, and much-needed, success.

Gary R. Whittaker is in the Department of Microbiology and Immunology, and in the Master of Public Health program, Cornell University, Ithaca, New York 14853, USA. Susan Daniel is in the Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, USA.
e-mail: grw7@cornell.edu

1. Pinto, D. *et al. Nature* **583**, 290–295 (2020).
2. Robbiani, D. F. *et al. Nature* https://doi.org/10.1038/s41586-020-2456-9 (2020).
3. Casadevall, A. & Pirofski, L. A. *J. Clin. Invest.* **130**, 1545–1548 (2020).
4. Watanabe, Y. *et al. Nature Commun.* **11**, 2688 (2020).
5. ter Meulen, J. *et al. PLoS Med.* **3**, e237 (2006).
6. Shi, R. *et al. Nature* https://doi.org/10.1038/s41586-020-2381-y (2020).
7. Ju, B. *et al. Nature* https://doi.org/10.1038/s41586-020-2380-z (2020).
8. Yuan, M. *et al. Science* **368**, 630–633 (2020).
9. Baum, A. *et al. Science* https://doi.org/10.1126/science.abd0831 (2020).
10. Rogers, T. F. *et al. Science* https://doi.org/10.1126/science.abc7520 (2020).
11. Brouwer, P. J. M. *et al. Science* https://doi.org/10.1126/science.abc5902 (2020).
12. Hansen, J. *et al. Science* https://doi.org/10.1126/science.abd0827 (2020).
13. Wec, A. Z. *et al. Science* https://doi.org/10.1126/science.abc7424 (2020).
14. Kelley, B. *Nature Biotechnol.* **38**, 540–545 (2020).
15. Furuyama, W. *et al. Sci. Rep.* **6**, 20514 (2016).
16. Saphire, E. O. *et al. Cell* **174**, 938–952 (2018).
17. Mulangu, S. *et al. N. Engl. J. Med.* **381**, 2293–2303 (2019).

Environmental science

Atmospheric CO_2 removed by rock weathering

Johannes Lehmann & Angela Possinger

Large-scale removal of carbon dioxide from the atmosphere might be achieved through enhanced rock weathering. It now seems that this approach is as promising as other strategies, in terms of cost and CO_2-removal potential. **See p.242**

Achieving targets for mitigating global warming will require the large-scale withdrawal of carbon dioxide from the atmosphere. On page 242, Beerling *et al.*[1] report that enhanced rock weathering in soils has substantial technical and economic potential as a global strategy for removing atmospheric CO_2. When crushed basalt or other silicate material is added to soil, it slowly dissolves and reacts with CO_2 to form carbonates. These either remain in the soil or move towards the oceans. The authors argue that this method would enable between 0.5 billion and 2 billion tonnes of CO_2 to be removed from the atmosphere each year. This rate is similar to that of other land-based approaches[2], such as the accrual of organic carbon in soil, carbon capture and sequestration in geological formations, and the addition of biochar (a carbon-rich material) to soil.

Beerling and colleagues find that removing atmospheric CO_2 through enhanced rock weathering would cost, on average, US\$160–190 per tonne of CO_2 in the United States, Canada and Europe, and \$55–120 per tonne of CO_2 in China, India, Mexico, Indonesia and Brazil. Furthermore, the authors report that China, the United States and India — the three largest emitters of CO_2 from fossil-fuel use — have the highest potential for CO_2 removal using this method. However, they also note that the application of silicate material to soil (Fig. 1) requires careful assessment of the risks, such as the possible release of metals and persistent organic compounds (compounds resistant to environmental degradation).

Despite the enthusiasm the authors' findings might generate, it is crucial to stress that, even under optimistic assumptions, enhanced rock weathering will sequester only some of the annual global carbon emissions from fossil-fuel use. Therefore, reducing these emissions should still be the top priority for averting dangerous climate change. But, as Beerling *et al.* note, any approach is insufficient alone, and should be considered as part of a portfolio of options.

Several other land-based carbon-sequestration techniques rely on soils. However, inorganic-carbon sequestration by rock weathering is fundamentally different from organic-carbon sequestration. The latter relies on photosynthesis by plants to remove CO_2 from the atmosphere, and on soils to retain the plant carbon, mostly in the form of microbial remains. In the future, therefore, scientists should pay closer attention to what they mean by 'carbon sequestration' — is it inorganic or organic?

The sequestration of atmospheric CO_2 through enhanced rock weathering shares some of the principal appeal, but also the challenges, of organic-carbon sequestration. The fact that crop production benefits is certainly a key asset of both methods. In the

ILSA B KANTOLA

Figure 1 | Application of silicate material to cropland. Beerling *et al.*[1] demonstrate that enhanced rock weathering, achieved by adding crushed basalt or other silicate material to soil, is an effective strategy for removing carbon dioxide from the atmosphere.

case of enhanced rock weathering, the added rock contains essential plant nutrients, such as calcium and magnesium, as well as potassium and micronutrients that promote crop production in several ways. We would go even further than the authors do, to claim that these nutrients are currently insufficiently supplied in agriculture.

Increasing soil pH alone would substantially boost crop yields in many regions of the world, because it is possible that low pH constrains crop production on more than 200 million hectares of arable and orchard soils[3]. This area is equivalent to about 20% of the total extent of these soils (967 million hectares; see go.nature.com/31rcajd). Consequently, on a global scale, acidity is the most important soil constraint for agriculture[4]. However, there have been no detailed multi-regional analyses of the difference in crop yield between low-pH and optimum-pH soils, and such investigations would benefit the study of synergies between carbon-sequestration methods. The proposed rock additions could conceivably mitigate the low use and supply shortages of agricultural limestone in several regions[5]. Furthermore, calcium improves root growth in acidic subsurface soil[6], with crucial knock-on effects through greater water uptake by plant roots.

Co-deployment of enhanced rock weathering with other soil-based sequestration approaches might both reduce limitations and maximize synergies[7]. Beerling and colleagues' study hints at some of these opportunities and at constraints that have procedural and soil-biogeochemical aspects. Greater crop growth will increase the input of crop residue (the materials from crops that are left in a field after harvesting) to the soil, and thereby enhance the accrual of organic carbon. However, the possibility that interactions between calcium and organic matter impede the return of CO_2 to the atmosphere has been sparsely explored, and there is little information on the effects of magnesium. In principle, calcium can reduce the decomposition of organic matter by facilitating adsorption to clay, inclusion in carbonates or aggregation[8]. But the indirect effects of calcium through changes in microbial ecology or interactions with organic compounds, rather than interactions only between organic compounds and clay minerals, are rarely studied.

If the synergy becomes a trade-off between organic-carbon sequestration and crop production, the organic-carbon content of soil could decrease, threatening the livelihoods of farmers, and even food security. Any carbon sequestration involving soils is a formidable challenge to incentivize, predict and monitor[9], because the sequestration technologies must be used on vast areas of land that are operated by hundreds of millions of farmers. Inevitably, there will be individual cases in which positive-yield projections are not met or crop yields even decrease, where incentives fail to persuade farmers, or where supply chains break down. But scientists should not be deterred from evaluating such technologies, and should instead accept that farmers need to be in the driving seat in adapting soil management to meet their specific site and crop-production goals. A concerted global effort will be required to develop site-specific optimization through farmer-centred research.

Fertilizer distribution networks are common in many parts of the world. But even where these networks are in place, success in the adoption of enhanced rock weathering might not rely on its crop-production benefits alone. We posit that carbon markets are required, and that it would be helpful if they incentivized socially and environmentally sound implementation[10]. For technologies to be eligible, it must be shown that they provide extra incentives for adoption (additionality), beyond what increased soil fertility would deliver. We emphasize that implementation of enhanced rock weathering and other soil-based carbon sequestration must consider equitable and financially sound incentives for farmers that overcome challenges of additionality, among others[10], in a proactive way.

"China, the United States and India have the highest potential for CO_2 removal using this method."

Consequently, the main lesson here might be that several of the major potential technologies for removing atmospheric CO_2 could generate substantial benefits for food production, and are centred around managing soils. Farmers must be fully behind such a global effort or it will fail. Scientists might need to recognize that climate-change mitigation is not a sufficient incentive on its own, and that benefits to crop growth will need to be prioritized, as will financial incentives. Such an approach of financially supporting soil health and crop production could emerge as our best near-term solution to the problem of removing CO_2 from the atmosphere.

Johannes Lehmann is in the Department of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, Ithaca, New York 14853, USA. **Angela Possinger** is in the Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, Virginia 24061, USA.
e-mail: cl273@cornell.edu

1. Beerling, D. J. *et al. Nature* **583**, 242–248 (2020).
2. Pacala, S. *et al. Negative Emissions Technologies and Reliable Sequestration: A Research Agenda* (National Academy of Sciences, 2018).
3. von Uexküll, H. R. & Mutert, E. *Plant Soil* **171**, 1–15 (1995).
4. Wood, S., Sebastian, K. & Scherr, S. J. *Pilot Analysis of Global Ecosystems: Agroecosystems* (World Resources Inst., 2000).
5. Fabregas, R., Kremer, M., Robinson, J. & Schilbach, F. *Evaluating Agricultural Information Dissemination in Western Kenya* (International Initiative for Impact Evaluation, 2017).
6. Lynch, J. P. & Wojciechowski, T. *J. Exp. Bot.* **66**, 2199–2210 (2015).
7. Amann, T. & Hartmann, J. *Biogeosciences* **16**, 2949–2960 (2019).
8. Rowley, M. C., Grand, S. & Verrecchia, É. P. *Biogeochemistry* **137**, 27–49 (2018).
9. Bradford, M. A. *et al. Nature Sustain.* **2**, 1070–1072 (2019).
10. Cavanagh, C. J., Vedeld, P. O., Petursson, J. G. & Chemarum, A. K. *J. Peasant Stud.* https://doi.org/10.1080/03066150.2019.1707812 (2020).

https://doi.org/10.1038/d41586-020-01916-2

News & views

Immunology

An antiviral response beyond immune cells

Tomás Gomes & Sarah A. Teichmann

Fibroblast, epithelial and endothelial cells are more than just the scaffold of an organ – it emerges that they communicate with immune cells and are primed to launch organ-specific gene-expression programs for antiviral defence.

Immune-system responses to disease-causing agents rely on a complex web of interactions between immune cells that are underpinned by robust regulatory mechanisms. Most of our understanding of the immune system revolves around these cells, yet cells generally thought of as having a mainly structural role can also respond to invading organisms. Writing in *Nature*, Krausgruber *et al.*[1] report a multi-organ examination of gene-expression programs for such structural cells in mice, revealing the roles of these cells in signalling networks used for defence purposes. The authors found that the response of structural cells to external invaders is regulated and tailored to the particular organ in question.

Structural cells, such as fibroblasts and endothelial and epithelial cells (Fig. 1), are present in most organs and provide more than just support[2,3]. Fibroblasts form part of the connective tissue and help to maintain the extracellular matrix material that surrounds cells. Endothelial cells line the interior of vessels such as blood vessels and, along with epithelial cells, which are present on the surface of organs, can be involved in responses to infection, either directly or through interactions with immune cells[3].

To understand the role of these three types of cell in immune responses, Krausgruber and colleagues isolated them from 12 different tissues in healthy mice. The authors used RNA sequencing to determine the genes expressed by the cells, and searched for known immune-associated genes. Krausgruber *et al.* also characterized the cells' chromatin – the complex of DNA and protein in the nucleus – to pinpoint genomic regions that were poised to start gene expression. This was done using a method called ATAC-seq to determine genome-wide 'open' chromatin accessibility, and the authors identified active promoter regions by tracking a type of modification called H3K4me2 on the DNA-binding histone 3 protein. Together, these methods opened a window on the transcriptional regulatory circuits that govern the identity and function of these cells.

Although the three cell types can be defined by the expression of genes corresponding to specific marker proteins found on the cell surfaces, the three cellular lineages also presented features that were characteristic of their local organ environment. Across the genome, the data sets for gene expression, open chromatin and active promoters indicated that the different cell types in an organ were more similar to each other than was a given cell type to the same cell type in different organs. This is a crucial observation that provides a foundation for future studies on the specific role that structural cells have in the function of each organ.

The authors searched the gene-expression data of structural cells to see which receptors and ligand molecules they expressed, and then matched the cells to possible interaction partners by mining previously published RNA-sequencing data for immune cells. They then assembled a computationally derived network that unveils possible cell-type- and organ-specific interactions involving structural and immune cells, and defines the baseline for

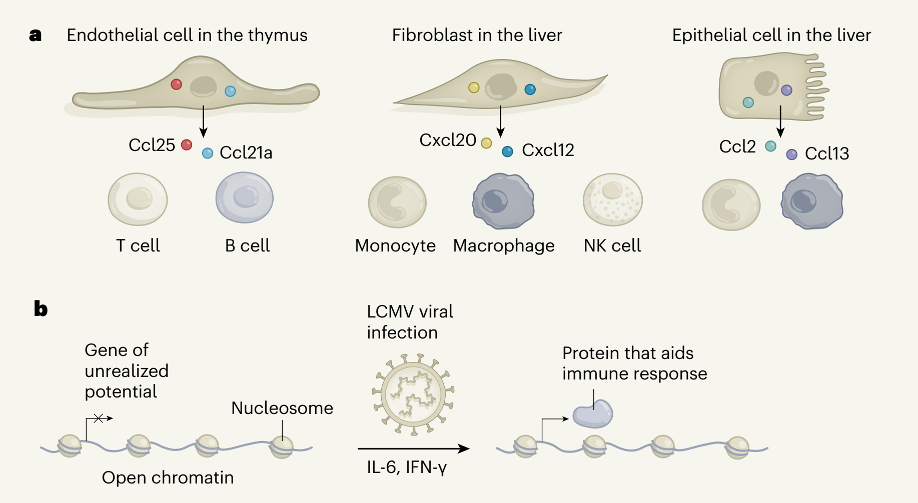


Figure 1 | Structural cells are poised for organ-specific defence responses. **a**, Krausgruber *et al.*[1] analysed three cell types – fibroblasts, and endothelial and epithelial cells – that are usually considered to have a structural role in organs. They found that, in mice, these cells signal to and interact with cells of the immune system (such as T cells, B cells, monocytes, macrophages and NK cells) to provide organ-specific defence responses. The authors report that structural cells express genes encoding chemokine proteins (for the examples given, the chemokines were Ccl25, Ccl21a, Cxcl10, Cxcl12, Ccl2 and Ccl13) that can attract immune cells. Structural cells also express other genes encoding ligands and receptors (not shown) that might aid communication with immune cells. The molecular interaction patterns identified were usually unique to each organ. **b**, Krausgruber *et al.* used RNA sequencing to profile gene expression in structural cells, and also assessed the state of chromatin (DNA wrapped around structures called nucleosomes) in the cells. Some genes were poised for expression – they had chromatin in an open state, and the authors described these genes as having unrealized potential. After infection with lymphocytic choriomeningitis virus (LCMV), these genes were expressed in a process that was often aided by the cytokine proteins IL-6 and IFN-γ (possibly secreted by immune cells). These genes were activated in a cell-type- and organ-specific manner, and constituted a key part of the early response of structural cells to infection.

routine interactions between immune cells and structural cells, from which further cellular crosstalk would develop on infection.

To understand more about how structural cells might prepare to trigger a gene-expression program for defence purposes, the authors assessed their gene-expression data together with the chromatin-accessibility profiles of the corresponding gene promoters (DNA sequences that aid gene expression). An open chromatin region encompassing a gene's promoter is known to be a reliable indicator of expression of the gene[4]. The authors used these combined data to look for outliers — genes that had an open accessible promoter but low levels of expression, on the assumption that such genes have what Krausgruber and colleagues describe as unrealized potential. This indicates genes that are probably poised for a rapid response when infection occurs. The approach highlighted a group of genes encoding a substantial number of immune-associated proteins, and examples of these were most evident in structural cells from the skin, liver and spleen. These genes are worthy of further study that focuses on how the structural cells that express them respond to infection and protect the organ that is their home.

The authors confirmed that they had indeed identified genes poised for a role in an immune response by infecting mice with lymphocytic choriomeningitis virus (LCMV) and then monitoring gene expression by RNA sequencing of structural cells. LCMV is a well-studied virus that affects most organs, and this allowed Krausgruber and colleagues to distinguish organ-specific from global defence responses. Eight days after infection, up to 57.9% of the genes of unrealized potential had been activated in structural cells, with notably high responses in fibroblasts and endothelial cells in the liver, spleen, lungs and large intestine.

Furthermore, the authors found that an antiviral response was evident in these gene-expression profiles. When infected and non-infected animals were compared, the infected animals had higher levels of expression of transcription factors and immune-associated signalling proteins called cytokines that are involved in pathways associated with expression of the antiviral protein interferon. In response to the viral infection, structural cells also expressed small proteins called chemokines that attract immune cells. This was a surprise, because chemokine secretion has been mainly associated with immune cells. The authors propose that their predicted interaction network between immune cells and structural cells is altered on LCMV infection, and suggest that, on infection, structural cells in various organs increase interactions with immune cells such as monocytes, macrophages and B cells.

To dissect the effects of signalling in response to LCMV infection, the authors injected individual cytokines, of types detected in the antiviral response, into the bloodstream of mice that did not have an LCMV infection. Krausgruber *et al.* then sequenced the RNA in structural cells from the organs with the greatest previously observed response to LCMV. They found that gene-expression changes were more evident in fibroblasts and endothelial cells than in epithelial cells. Dissecting the gene-expression response to each cytokine revealed the portion of the antiviral program that it controls. Among other interactions, this revealed that the cytokines IL-6 and IFN-γ, possibly produced *in vivo* by immune cells, are responsible for eliciting much of the antiviral response of spleen endothelial cells by driving the expression of genes with unrealized potential.

Although gene-expression programs involved in the immune response have been reported previously for some structural cells, Krausgruber and colleagues' work underscores these cells' decisive role in coordinating organ-specific and organism-wide immune responses. It also indicates how functionally relevant candidate genes can be pinpointed using a combination of cell-communication networks and analysis of chromatin-mediated regulation. One of the ultimate goals of this research field could be to develop cell-type-targeted therapies that modulate immune responses. This could greatly benefit cancer research, for example, because cancer-associated fibroblasts have a role in promoting tumour progression[5].

Future studies will probably focus on the defence responses of other types and subtypes of human cells in studies linked to the Human Cell Atlas initiative[6], which is generating detailed molecular profiles for all human cells to fully describe cell-type diversity. Single-cell approaches could assist in profiling the RNA transcripts in all cell types and states of entire organs, in steady-state and post-stimulus scenarios. The use of a new method called spatial transcriptomics (which monitors gene expression in intact tissue sections rather than in dissociated cells), together with information about chromatin status, could disentangle the entire cellular chain of events, from the detection of infection to the defence response and immune-cell recruitment, and then finally to the removal of the infectious agent. By profiling structural cells in different mouse organs, Krausgruber *et al.* have unlocked a trove of knowledge about antiviral defences, which might be relevant to other species and facilitate new ways to target human diseases.

Tomás Gomes and **Sarah A. Teichmann** are at the Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. **S.A.T.** is also at the Cavendish Laboratory, Department of Physics, University of Cambridge, UK.
e-mails: tpcg@sanger.ac.uk; st9@sanger.ac.uk

1. Krausgruber, T. *et al. Nature* https://doi.org/10.1038/s41586-020-2424-4 (2020).
2. Van Linthout, S., Miteva, K. & Tschöpe, C. *Cardiovasc. Res.* **102**, 258–269 (2014).
3. Young, M. R. *Cancer Immunol. Immunother.* **61**, 1609–1616 (2012).
4. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. *Nature Rev. Genet.* **20**, 207–220 (2019).
5. Cirri, P. & Chiarugi, P. *Am. J. Cancer Res.* **1**, 482–497 (2011).
6. Regev, A. *et al. eLife* **6**, e27041 (2017).

https://doi.org/10.1038/d41586-020-01815-6

News & views

Cancer evolution

Strands of evidence

Trevor A. Graham & Sarah E. McClelland

DNA damage can cause mutations due to failure of DNA repair and errors during DNA replication. Tracking the strand of the DNA double helix on which damage occurs has shed light on processes that affect tumour evolution.

How a cancer evolves and how mutations are generated are highly intertwined processes, and both are nearly impossible to observe directly. Instead, we are usually restricted to making inferences about them using data from a single snapshot in time after a cancer has formed. Writing in *Nature*, Aitken *et al.*[1] show that, for a cell that has undergone DNA damage, such a snapshot provides remarkably rich information when the two DNA strands that form the double helix are considered independently.

DNA resembles a ladder, with the two 'side rails' often called, respectively, the Watson and Crick strands. These are fused together by 'rungs' of two complementary nucleotide base pairs: either cytosine (C) paired with guanine (G) or adenine (A) paired with thymine (T). When a cell divides, each daughter cell inherits either the Watson or Crick strand from the parent; this provides a template from which the other, complementary strand is replicated. Damage to a base can trigger a repair process, but if repair is not swift enough, the damaged base might be mispaired with an incorrect base during DNA replication. At the next round of cell division, when a daughter cell with such a mispaired base prepares to divide, the base complementary to the mispaired base will be added to the newly synthesized strand. This leads to a double-stranded mutation at the base pair corresponding to the original damaged base (Fig. 1).

Standard practice for genome sequencing is to consider mutations without paying attention to which of the strands received the original damage. However, when a chemical change occurs that damages a base, creating a site referred to as a lesion, this lesion is on only one of the two DNA strands of the affected base pair. Aitken and colleagues had the insight to see that, because the 'parental' Watson and Crick strands of an original cell that underwent DNA damage are separated into different daughter cells, when the cell divides, two cell lineages can be tracked individually by following the unique pattern of mutations that lesions on each of the parental strands generates.

To induce DNA lesions, Aitken and colleagues gave mice a large dose of the carcinogenic molecule diethylnitrosamine. This treatment predominantly caused DNA lesions at T bases in liver cells, ultimately leading to tumour growth. When the authors examined the pattern of diethylnitrosamine-induced mutations along the genome of each tumour, they found long stretches of the genome that, compared with the original, unmutated genome, were highly enriched for mutations in which T was mutated to any other base (N). These mutations were derived from T lesions on the DNA strand (let's call it the Crick strand) that was inherited by the daughter cell, and its cellular descendants, that went on to form the tumour. Lesions on either strand can generate mutations, but it might be the case that lesions on only one of the parental strands generates tumour-promoting mutations and hence

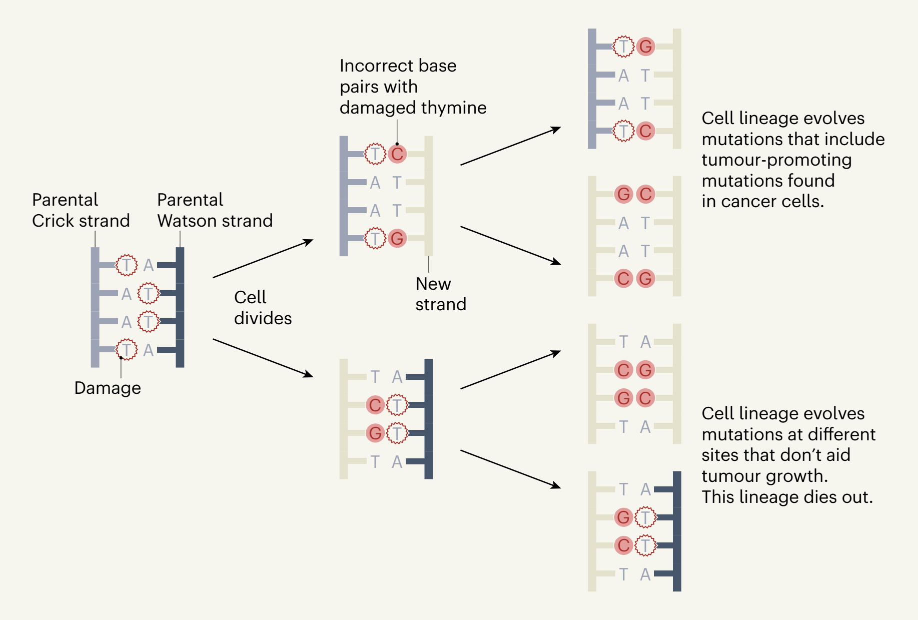


Figure 1 | Tracking the connection between damage to individual DNA strands and tumour mutations. Aitken *et al.*[1] analysed mutation patterns in the cells of mice that received a carcinogenic molecule called diethylnitrosamine, which damages thymine (T) nuclotide bases. Damage is indicated by red circles. The cell originally exposed to diethylnitrosamine has two DNA strands, named here as the parental Crick and Watson strands. The correct pairing of bases is either adenine (A) paired with T or guanine (G) paired with cytosine (C). When the original cell divides, an incorrect base (shown in pink) can mispair with a damaged T base. Then, as those two daughter cells divide, the incorrect base will pair with the complementary matching base (for example, a mispaired C pairs with a G on the newly made strand), which results in both DNA strands having a mutation at that particular base pair. The original cell division generates two cell lineages that have mutations arising from damage to either the parental Crick or parental Watson strand, respectively. Each of these lineages has mutations at distinct base-pair locations. As the cells continue to divide, mispairing opposite unrepaired damaged T bases continues, producing different mutations at the same base-pair position and generating genetic diversity. Only cell lineages with mutations that aid tumour growth will be found in a tumour that the mouse subsequently develops. This example shows one possible scenario, in which mutations arising as a consequence of damage only to the parental Crick strand contribute to cancer growth.

only one of the two daughter lineages forms a tumour. In the example shown in Figure 1, the lesions on T bases on the corresponding Watson parental strand received by the other daughter cell did not lead to the formation of tumour-promoting mutations, and this cellular lineage therefore did not contribute to the tumour.

The authors realized that the pattern of base-pair locations that had T-to-N mutations enabled them to pinpoint the individual Watson or Crick strand that had served as the template strand for the first cell in the tumour. This template strand carried diethylnitrosamine-induced lesions mainly at T bases, allowing the fate of each strand to be tracked individually through subsequent cell divisions (Fig. 1). These individual 'strands of evidence' provide remarkable information about the process of mutation and tumour evolution.

In gene expression, DNA is transcribed to produce RNA, and DNA lesions can be repaired by a process called transcription-coupled repair[2]. Aitken and colleagues observed that transcription-coupled repair occurred preferentially on the strand being transcribed, as opposed to the complementary strand, and that higher levels of transcription were associated with an increased frequency of repair.

The authors found that failure to repair a DNA lesion over successive cell cycles, an interesting observation in itself, provided an unexpected source of genetic diversity. Each round of DNA replication on a lesion-containing strand could lead to the incorporation of a different 'wrong', mispaired base opposite the lesion site in the newly synthesized strand. If this happened, it caused further, distinct mutations at the same genomic position, generating cells in the tumour each with different mutations of the same base pair.

Observing recurrent mutations at the same genomic site could be taken as evidence of convergent evolution, in which multiple individual mutational events at that base-pair site are all positively selected for during tumour growth. Instead, Aitken and colleagues' findings indicate that recurrent mutations could result from lesion-bearing DNA strands being used as templates for DNA replication over multiple rounds of cell division.

Intriguingly, strand tracing also provides a window on the selection of mutations associated with cancer. When a cell divides, a daughter cell should inherit, at random, either DNA strand. However, when the authors tracked the prevalence of sequences corresponding to inheritance of the parental Watson or Crick strand of a particular chromosome, they noticed that the tumours contained one of these two strands more often than would be expected by chance.

The authors' explanation for such preferential strand retention is that it occurred because the retained strand contained a diethylnitrosamine-induced mutation in a gene that is important for tumour growth. Aitken *et al.* identified three potential tumour-promoting genes in this way, all of which are known to be crucial for the growth of liver tumours. Strand-by-strand analysis might be an unexpectedly useful tool for probing the tumour-promoting contribution of non-protein-coding regions of the genome, because selection can be detected without needing to know the background mutation rate — the problem of determining this rate has posed a challenge for methods previously used to study these regions[3].

It might be expected that because there is strand-biased prevalence of mutations corresponding to diethylnitrosamine-induced lesions, a chromosome should be enriched for T-to-N mutations along the entire length of its DNA strand. Instead, the authors found that, for a single chromosome, the enrichment of such mutations sometimes switched over to the other strand (and could be observed as A-to-N mutations). They propose that this provides evidence of sites of a DNA-repair process called homologous recombination, in which DNA strands from a chromosome exchange with identical DNA sequences in a cell that is gearing up to divide, during an event called sister-chromatid exchange.

Sister-chromatid exchange is usually an elusive process to monitor because it involves, in theory, an exchange between two identical DNA sequences. However, because Aitken and colleagues could track individual strands through mutation patterns, they could detect evidence of such events. Interestingly, a higher frequency of these exchange events tracked with a higher diethylnitrosamine-induced mutation burden, suggesting that diethylnitrosamine-induced damage to DNA might prompt sister-chromatid exchange[4]. Aitken *et al.* report that presumptive exchange events tended to occur in regions of the genome that were associated with lower-than-average gene expression and later replication, compared with other regions, during the cell cycle. Interestingly, these features are hallmarks of 'fragile sites' — genomic regions susceptible to damage during DNA replication that are known to be prone to sister-chromatid exchange[5,6].

The most tantalizing question raised by this study is whether the DNA strand (the original Watson or Crick strand) in which damage first occurred can be tracked in human cancers. People are usually exposed to mutation-causing agents, such as those in cigarette smoke, for long periods of time, making it probable that DNA lesions are continually being induced. Consequently, the mutational signal arising from an individual DNA strand would probably be obscured. However, when Aitken and colleagues assessed data already available from human cancers, they found that, in rare cases of sudden, acute exposure to a mutagenic agent, most clearly observed for aristolochic acid exposure (which leads to liver, kidney and bile-duct cancer), mutational signals could be traced back to identify the strand originally damaged.

Chemotherapy also provides an example of acute exposure to mutation-causing agents, and can give rise to distinctive mutational signatures[7]. It will be interesting to see what can be learnt by applying Aitken and colleagues' methods to analysing chemotherapy-treated cancers. Other examples of exposure to mutation-causing agents, such as acute radiation exposure or sunburn, might also be worth analysing using a strand-by-strand approach. Furthermore, such analysis could clarify the timeline of exposure to a mutation-causing agent: a single, large exposure might generate a mutational signal that could be assigned to individual DNA strands, whereas repeated exposures might cause a progressively less distinct signal. Similarly, analysing individual strands might provide insight into the rate of lesion-repair processes, and offer a new means of studying defects in DNA-repair processes in cancer, such as defective-mismatch repair.

Tracking individual DNA strands is a reductive approach offering a powerful way to study DNA replication and repair processes that have been challenging to observe. Aitken and colleagues' study shows the potential of this method for tackling the complexities of cancer — it seems that the maxim that the whole is greater than the sum of the parts does not apply to individual DNA strands.

Trevor A. Graham and **Sarah E. McClelland** are at the Centre for Cancer Genomics and Computational Biology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK.
e-mails: t.graham@qmul.ac.uk; s.mcclelland@qmul.ac.uk

1. Aitken, S. J. *et al. Nature* https://doi.org/10.1038/s41586-020-2435-1 (2020).
2. Strick, T. R. & Portman, J. R. *J. Mol. Biol.* **431**, 4093–4102 (2019).
3. Rheinbay, E. *et al. Nature* **578**, 102–111 (2020).
4. Verna, L., Whysner, J. & Williams, G. M. *Pharmacol. Ther.* **71**, 57–81 (1996).
5. Glover, T. W. & Stein, C. K. *Am. J. Hum. Genet.* **41**, 882–890 (1987).
6. Waisertreiger, I., Popovich, K., Block, M., Anderson, K. R. & Barlow, J. H. *Oncogene* **39**, 1260–1272 (2020).
7. Pich, O. *et al. Nature Genet.* **51**, 1732–1740 (2019).

Article

Asymmetric mass ratios for bright double neutron-star mergers

https://doi.org/10.1038/s41586-020-2439-x

Received: 19 December 2019

Accepted: 14 April 2020

R. D. Ferdman[1✉], P. C. C. Freire[2], B. B. P. Perera[3], N. Pol[4,5], F. Camilo[6], S. Chatterjee[7,8], J. M. Cordes[7,8], F. Crawford[9], J. W. T. Hessels[10,11], V. M. Kaspi[12,13], M. A. McLaughlin[4,5], E. Parent[12,13], I. H. Stairs[14] & J. van Leeuwen[11]

The discovery of a radioactively powered kilonova associated with the binary neutron-star merger GW170817 remains the only confirmed electromagnetic counterpart to a gravitational-wave event[1,2]. Observations of the late-time electromagnetic emission, however, do not agree with the expectations from standard neutron-star merger models. Although the large measured ejecta mass[3,4] could be explained by a progenitor system that is asymmetric in terms of the stellar component masses (that is, with a mass ratio q of 0.7 to 0.8)[5], the known Galactic population of merging double neutron-star systems (that is, those that will coalesce within billions of years or less) has until now consisted only of nearly equal-mass ($q > 0.9$) binaries[6]. The pulsar PSR J1913+1102 is a double system in a five-hour, low-eccentricity (0.09) orbit, with an orbital separation of 1.8 solar radii[7], and the two neutron stars are predicted to coalesce in 470^{+12}_{-11} million years owing to gravitational-wave emission. Here we report that the masses of the pulsar and the companion neutron star, as measured by a dedicated pulsar timing campaign, are 1.62 ± 0.03 and 1.27 ± 0.03 solar masses, respectively. With a measured mass ratio of $q = 0.78 \pm 0.03$, this is the most asymmetric merging system reported so far. On the basis of this detection, our population synthesis analysis implies that such asymmetric binaries represent between 2 and 30 per cent (90 per cent confidence) of the total population of merging binaries. The coalescence of a member of this population offers a possible explanation for the anomalous properties of GW170817, including the observed kilonova emission from that event.

Since its discovery[7] in 2012, we have been regularly monitoring the double neutron star (DNS) PSR J1913+1102 with the Arecibo radio telescope. Our observations have used the Mock Spectrometer and the Puerto Rico Ultimate Pulsar Processing Instrument (PUPPI) to coherently remove dispersive smearing from the pulsar signal, caused by the interstellar free-electron plasma along the line of sight to the pulsar. We analysed data from this pulsar using standard pulse timing techniques (see Methods).

With a spin period of 27 ms, PSR J1913+1102 was probably the first-formed neutron star in this binary system; it subsequently gained angular momentum via accretion of matter from the progenitor to the second neutron star[7]. The timing of the pulsar has allowed a precise measurement of the rate of advance of the periastron, which is $\dot{\omega} = (5.6501 \pm 0.0007)^{\circ}\ \mathrm{yr}^{-1}$. In addition, we have now determined two more post-Keplerian parameters: the first is the Einstein delay ($\gamma = 0.471 \pm 0.015$ ms), which describes the effect of gravitational redshift and relativistic time dilation due to the varying orbital velocity and proximity of the neutron stars to one another during their orbits. The second is the decay of the orbital period caused by the emission of gravitational waves ($\dot{P}_{\mathrm{b}} = (-4.8 \pm 0.3) \times 10^{-13}\ \mathrm{s\ s}^{-1}$).

In Fig. 1, we show the general-relativistic mass constraints corresponding to each measured post-Keplerian parameter. From $\dot{\omega}$, the total system mass is $(2.8887 \pm 0.0006)M_{\odot}$ ($M_{\odot}$, solar mass), making PSR J1913+1102 the most massive among known DNS systems (by a 2% margin). By combining $\dot{\omega}$ and γ, we obtain the individual neutron-star masses $m_{\mathrm{p}} = (1.62 \pm 0.03)M_{\odot}$ and $m_{\mathrm{c}} = (1.27 \pm 0.03)M_{\odot}$ (unless otherwise stated, uncertainties denote 68% confidence) for the pulsar and the companion, respectively, which give a mass ratio of $q = m_{\mathrm{c}}/m_{\mathrm{p}} = 0.78 \pm 0.03$. The observed $\dot{P}_{\mathrm{b}}$ is consistent with the general relativity prediction for these neutron-star masses; apart from confirming them, this effect provides a unique test of alternative gravitational theories that will be reported elsewhere (P.C.C.F. et al., manuscript in

---

[1] Faculty of Science, University of East Anglia, Norwich, UK. [2] Max-Planck-Institut für Radioastronomie, Bonn, Germany. [3] Arecibo Observatory, Arecibo, Puerto Rico. [4] Department of Physics and Astronomy, West Virginia University, Morgantown, WV, USA. [5] Center for Gravitational Waves and Cosmology, West Virginia University, Morgantown, WV, USA. [6] South African Radio Astronomy Observatory, Cape Town, South Africa. [7] Cornell Center for Astrophysics and Planetary Science, Cornell University, Ithaca, NY, USA. [8] Department of Astronomy, Cornell University, Ithaca, NY, USA. [9] Department of Physics and Astronomy, Franklin and Marshall College, Lancaster, PA, USA. [10] Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [11] ASTRON, Netherlands Institute for Radio Astronomy, Dwingeloo, The Netherlands. [12] Department of Physics, McGill University, Montreal, Quebec, Canada. [13] McGill Space Institute, McGill University, Montreal, Quebec, Canada. [14] Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada.
✉e-mail: r.ferdman@uea.ac.uk

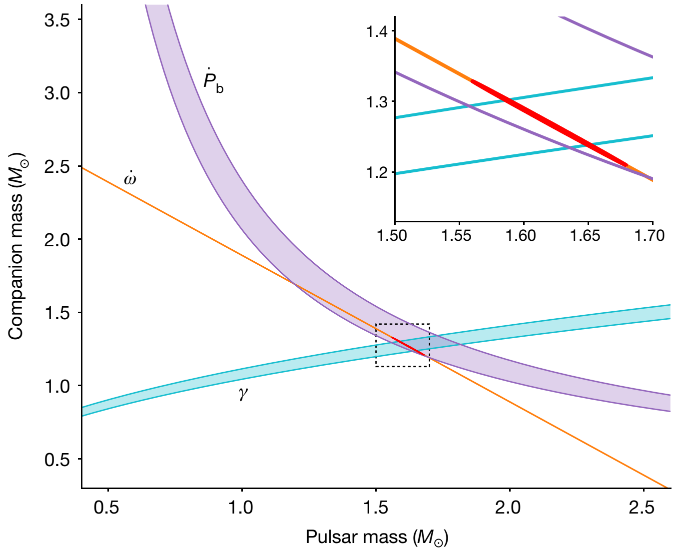


Fig. 1 | Pulsar mass–companion mass diagram for the PSR J1913+1102 system. Shaded regions bounded by solid curves represent 1σ mass constraints from each measured post-Keplerian parameter, derived in the context of general relativity. These are: the orbital precession rate ($\dot{\omega}$), the time dilation/gravitational redshift (γ) and the rate of orbital decay ($\dot{P}_b$). The inset shows a zoom-in of the dotted square region in the main plot, with the 3σ confidence region for the mass measurements shaded in red. The two most precisely measured parameters allow us to determine the individual masses of this system. Each additional post-Keplerian parameter measurement provides an independent consistency test of the predictions of general relativity.

preparation). Table 1 summarizes the best-fit model parameters for the PSR J1913+1102 system.

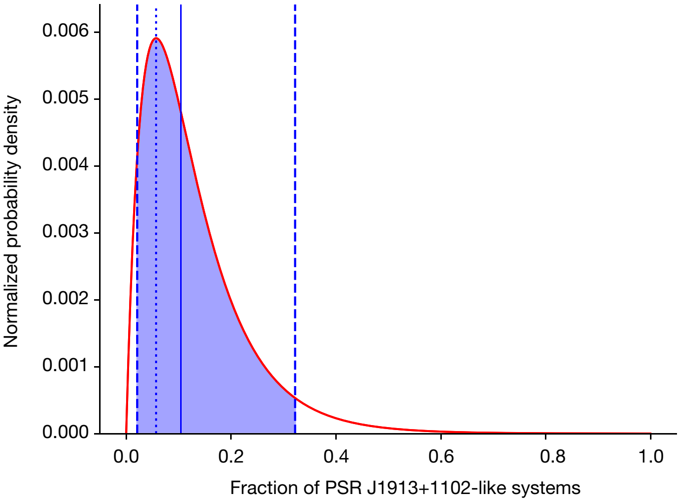


Fig. 2 | Probability density of the population of PSR J1913+1102 -like DNS systems in the Galaxy, as a fraction of the total number of DNSs that will merge within a Hubble time. We find this fraction to be $0.11^{+0.21}_{-0.09}$, where the uncertainty represents the 90% confidence interval (vertical dashed lines). The quoted value is the median of the distribution, shown on the plot as a solid vertical line, and the peak value of 0.06 is represented by a dotted vertical line. This implies that roughly 1 in 10 merging DNS systems are likely to have asymmetric component masses.

Table 1 | Measured and derived parameters for PSR J1913+1102

Parameter name	Value
Reference epoch (MJD)	57,504.0
Observing time span (MJD)	56,072–58,747
Number of arrival time measurements	2,541
Solar System ephemeris used	DE436
Root-mean-square timing residual	56 μs
Reduced χ^2 of timing fit	1.01
Right ascension, α (J2000)	19 h 13 min 29.05365(9) s
Declination, δ (J2000)	11° 02′ 05.7045(22)″
Proper motion in α	−3.0(5) mas yr^{-1}
Proper motion in δ	−8.7(1.0) mas yr^{-1}
Pulsar spin period, P	27.2850068680286(19) ms
Spin period derivative, $\dot{P}$	1.5672(7) × 10^{-19} s s^{-1}
Dispersion measure, DM	339.026(3) pc cm^{-3}
Orbital period, P_b	0.2062523345(2) d
Projected semi-major axis of the pulsar's orbit, x	1.754635(5) light s
Orbital eccentricity, e	0.089531(2)
Longitude of periastron, ω	283.7898(19)°
Epoch of periastron passage, T_0 (MJD)	57,504.5314530(10)
Total system mass, M	2.8887(6)M
Companion mass, M_c	1.27(3)M
Rate of periastron advance[a], $\dot{\omega}$	5.6501(7)° yr^{-1}
Einstein delay[a], γ	0.000471(15) s
Orbital period decay rate[a], $\dot{P}_b$	−4.8(3) × 10^{-13} s s^{-1}
Pulsar mass, M_p	1.62(3)
Mass ratio, q	0.78(3)
Orbital inclination angle, i	55.3°
Dispersion-derived distance[b], d	7.14 kpc
Two-dimensional systemic peculiar velocity[c], v_{LSR}	100(70) km s^{-1}
Surface magnetic flux density at the poles[d], B_0	2.1 × 10^9 G
Characteristic age[d], τ_c	2.8 Gyr
Time to coalescence, T_c	470^{+12}_{-11} Myr

Values in parentheses represent the 1σ (68% confidence) uncertainty on the last quoted digit. Unless otherwise noted, measured parameters were determined using the DDGR timing model[43,44], which assumes general relativity to be the correct theory of gravity. MJD, modified Julian date.

[a] Post-Keplerian orbital parameters were measured using the model-independent Damour and Deruelle timing model[43–45].

[b] Distance is derived using a model of the Galactic ionized electron density[46], with estimated uncertainty of 20%.

[c] Velocity is with respect to the local standard of rest, and thus peculiar to the binary system.

[d] The surface magnetic field and characteristic age are derived from the pulsar spin period and its derivative[47].

PSR J1913+1102 is part of a population of several very compact DNS binary systems with moderate orbital eccentricities (≲0.2) and low proper motions (for example, PSRs J0737−3039A/B[8], J1756−2251[9] and J1946+2052[10]). These imply an evolutionary path in which the second-formed neutron star is born as a result of an envelope-stripped helium star progenitor having undergone a supernova with very little mass loss and low natal kick[6,11], owing to either a rapid iron core collapse event or electron capture onto an oxygen–neon–magnesium core[12–14]. Either of these scenarios lead to a low-mass (≲1.3$M_\odot$) neutron star[7], which is confirmed by our measurements. Furthermore, we estimate the tangential component of the peculiar velocity of the system to be 100 ± 70 km s^{-1} (95% confidence); although the uncertainty is still large, this hints at a relatively low kick velocity arising from the second supernova compared to the overall pulsar population.

The PSR J1913+1102 mass ratio makes it the most asymmetric among the known DNS binaries that are expected to merge within a Hubble time, which otherwise have $q \gtrsim 0.9$. Considering all known DNS systems,

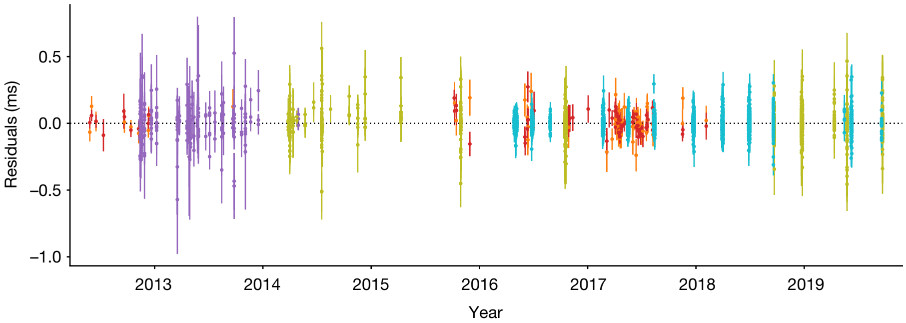


Fig. 3 | Post-fit timing residuals for PSR J1913+1102. These are obtained after including all best-fit parameters in the Damour and Deruelle General Relativity (DDGR) model[43,44] ephemeris for this pulsar. Each contributing instrument is represented by different colours: Mock spectrometer centred at 1,300 MHz (orange) and at 1,450 MHz (red); PUPPI centred at 1,400 MHz in incoherent mode (purple); PUPPI centred at 1,400 MHz in coherent-fold mode (yellow); and PUPPI centred at 2,350 MHz in coherent-fold mode (cyan). The latter provided a substantial improvement in data quality, as evidenced by the reduction in the weighted root-mean-square timing residuals to 48 μs, down from 72 μs at 1,400 MHz in coherent mode. The error bars shown reflect the 1σ (68%) uncertainties of each data point.

the only one having a similar mass asymmetry is PSR J0453+1559[15], with $q = 0.753 \pm 0.005$; however, its orbital period of 4.07 days implies a coalescence time about 100 times greater than the age of the Universe. By contrast, PSR J1913+1102 has an expected time to coalescence of 470^{+12}_{-11} Myr, which we determined from the orbital decay rate and other measured orbital elements.

There are currently nine confirmed compact DNS binaries that are predicted to merge within a Hubble time, for which precise neutron-star mass measurements have been made[6]. We performed a population synthesis analysis for these DNS systems, including PSR J1913+1102, using their individual properties and known masses (see also Methods). We found that PSR J1913+1102-like binaries represent 11^{+21}_{-9} % of merging DNS systems (Figs. 2, 3), where the quoted value is the median and the errors represent the 90% confidence intervals. This therefore establishes the existence of a population of asymmetric DNS systems that is sizable enough to potentially lead to the discovery of several corresponding merger events by ground-based gravitational-wave observatories such as LIGO/Virgo. As such, the discovery of PSR J1913+1102 provides evidence for the need to account for asymmetric coalescing DNS binaries—approximately one-tenth of events—in interpeting merger scenarios and the underpinning physics.

Observations of the electromagnetic counterparts to the GW170817 event have largely been related to a relatively large amount of ejecta due to the preceding DNS merger, with masses of the order of $0.05M_{\odot}$ (refs. [4,16]). This is at odds with standard models of DNS coalescence, which typically predict smaller ejecta mass by at least a factor of 5, primarily on the basis of the assumption of equal-mass (or near-equal-mass) progenitor DNS binary systems[5,17,18]. Until recently, this has been a reasonable assumption given the known DNS population[6]. It is plausible that the anomalously massive ejecta inferred from the observed late-time emission in the case of GW170817 may be explained with an equal-mass system, particularly if one includes a secular component to the ejecta[19–21]. The merger may also be explained by the invocation of various models to describe the observations. These include an off-axis jet from a short γ-ray burst[22,23]; a mildly relativistic wide-angle outflow that interacts with the dynamic ejecta[24,25]; a hypermassive neutron-star remnant of the merger acting as a spin-down energy source[26,27]; and a hierarchical triple system in which a disk is formed from a Roche lobe-filling outer star[28].

By contrast, numerical simulations have shown that high-asymmetry ($0.65 \lesssim q \lesssim 0.85$) systems will naturally produce larger tidal distortions during the merger phase and result in larger-mass, and therefore brighter, disks than those of roughly equal-mass systems[18,29–31]. The resulting tidal effects consistently produce sufficient neutron-rich ejecta to power a kilonova and result in an enhancement of r-process material[32]. A sufficiently unequal-mass binary that will merge within a Hubble time may therefore be responsible for events such as GW170817, particularly in the slow, $\sim 0.04M_{\odot}$ red component seen in the latter[4]. A substantial population of asymmetric DNSs such as PSR J1913+1102 would therefore lead to an enhanced detection rate of bright kilonovae. The electromagnetic counterparts to such events are therefore particularly important for understanding the Galactic heavy-element abundance[16,33,34].

Enhanced pre-coalescence tidal distortions due to asymmetric mergers may also allow certain neutron-star equation-of-state models to be ruled out through study of the gravitational-wave waveform[21,35–38]. Additionally, gravitational waves from merging DNSs have recently been used as distance indicators (so-called 'standard sirens'), enabling an independent probe of the Hubble constant, H_0, when combined with radial-velocity measurements of the electromagnetic counterparts[39]. Future asymmetric DNS mergers with similar electromagnetic counterparts to those of GW170817 would lead to a considerably more precise determination of H_0—about 15 suitable detections would provide a ~2% measurement[40]—and potentially provide the means by which to resolve the disagreement between H_0 as measured from the cosmic microwave background[41] and through local Universe analysis methods[42].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2439-x.

1. Abbott, B. et al. GW170817: observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.* **119**, 161101 (2017).
2. Abbott, B. P. et al. Multi-messenger observations of a binary neutron star merger. *Astrophys. J. Lett.* **848**, 12 (2017).
3. Abbott, B. P. et al. Estimating the contribution of dynamical ejecta in the kilonova associated with GW170817. *Astrophys. J. Lett.* **850**, 39 (2017).
4. Cowperthwaite, P. S. et al. The electromagnetic counterpart of the binary neutron star merger LIGO/Virgo GW170817. II. UV, optical, and near-infrared light curves and comparison to kilonova models. *Astrophys. J. Lett.* **848**, 17 (2017).
5. Pankow, C. On GW170817 and the Galactic binary neutron star population. *Astrophys. J.* **866**, 60 (2018).
6. Tauris, T. M. et al. Formation of double neutron star systems. *Astrophys. J.* **846**, 170 (2017).
7. Lazarus, P. et al. Einstein@Home discovery of a double neutron star binary in the PALFA survey. *Astrophys. J.* **831**, 150 (2016).
8. Kramer, M. et al. Tests of general relativity from timing the double pulsar. *Science* **314**, 97–102 (2006).
9. Ferdman, R. D. et al. PSR J1756−2251: a pulsar with a low-mass neutron star companion. *Mon. Not. R. Astron. Soc.* **443**, 2183–2196 (2014).
10. Stovall, K. et al. PALFA discovery of a highly relativistic double neutron star binary. *Astrophys. J. Lett.* **854**, 22 (2018).

11. Tauris, T. M. et al. Ultra-stripped type Ic supernovae from close binary evolution. *Astrophys. J. Lett*. **778**, 23 (2013).
12. Miyaji, S., Nomoto, K., Yokoi, K. & Sugimoto, D. Supernova triggered by electron captures. *Publ. Astron. Soc. Jpn*. **32**, 303–329 (1980).
13. Nomoto, K. Evolution of 8–10 solar mass stars toward electron capture supernovae. I – formation of electron-degenerate O + NE + MG cores. *Astrophys. J*. **277**, 791–805 (1984).
14. Podsiadlowski, P. et al. The double pulsar J0737–3039: testing the neutron star equation of state. *Mon. Not. R. Astron. Soc*. **361**, 1243–1249 (2005).
15. Martinez, J. G. et al. Pulsar J0453+1559: a double neutron star system with a large mass asymmetry. *Astrophys. J*. **812**, 143 (2015).
16. Kasen, D., Metzger, B., Barnes, J., Quataert, E. & Ramirez-Ruiz, E. Origin of the heavy elements in binary neutron-star mergers from a gravitational-wave event. *Nature* **551**, 80–84 (2017).
17. Hotokezaka, K. et al. Mass ejection from the merger of binary neutron stars. *Phys. Rev. D* **87**, 024001 (2013).
18. Radice, D. & Dai, L. Multimessenger parameter estimation of GW170817. *Eur. Phys. J. A* **55**, 50 (2019).
19. Siegel, D. M. & Metzger, B. D. Three-dimensional GRMHD simulations of neutrino-cooled accretion disks from neutron star mergers. *Astrophys. J*. **858**, 52 (2018).
20. Fernández, R., Tchekhovskoy, A., Quataert, E., Foucart, F. & Kasen, D. Long-term GRMHD simulations of neutron star merger accretion discs: implications for electromagnetic counterparts. *Mon. Not. R. Astron. Soc*. **482**, 3373–3393 (2018).
21. Radice, D. et al. Binary neutron star mergers: mass ejection, electromagnetic counterparts and nucleosynthesis. *Astrophys. J*. **869**, 130 (2018).
22. Troja, E. et al. The X-ray counterpart to the gravitational-wave event GW170817. *Nature* **551**, 71–74 (2017).
23. Haggard, D. et al. A deep Chandra X-ray study of neutron star coalescence GW170817. *Astrophys. J*. **848**, L25 (2017).
24. Mooley, K. P. et al. A mildly relativistic wide-angle outflow in the neutron-star merger event GW170817. *Nature* **554**, 207–210 (2018).
25. Piro, A. L. & Kollmeier, J. A. Evidence for cocoon emission from the early light curve of SSS17a. *Astrophys. J*. **855**, 103 (2018).
26. Metzger, B. D., Thompson, T. A. & Quataert, E. A magnetar origin for the kilonova ejecta in GW170817. *Astrophys. J*. **856**, 101 (2018).
27. Li, S.-Z., Liu, L.-D., Yu, Y.-W. & Zhang, B. What powered the optical transient AT2017gfo associated with GW170817? *Astrophys. J. Lett*. **861**, 12 (2018).
28. Chang, P. & Murray, N. GW170817: A neutron star merger in a mass-transferring triple system. *Mon. Not. R. Astron. Soc. Lett*. **474**, 12–16 (2018).
29. Shibata, M. & Taniguchi, K. Merger of binary neutron stars to a black hole: disk mass, short gamma-ray bursts, and quasinormal mode ringing. *Phys. Rev. D* **73**, 064027 (2006).
30. Rezzolla, L., Baiotti, L., Giacomazzo, B., Link, D. & Font, J. A. Accurate evolutions of unequal-mass neutron-star binaries: properties of the torus and short GRB engines. *Class. Quantum Gravity* **27**, 114105 (2010).
31. Dietrich, T. & Ujevic, M. Modeling dynamical ejecta from binary neutron star mergers and implications for electromagnetic counterparts. *Class. Quantum Gravity* **34**, 105014 (2017).
32. Lehner, L. et al. Unequal mass binary neutron star mergers and multimessenger signals. *Class. Quantum Gravity* **33**, 184002 (2016).
33. Tanvir, N. R. et al. A ‘kilonova’ associated with the short-duration γ-ray burst GRB 130603B. *Nature* **500**, 547–549 (2013).
34. Just, O., Bauswein, A., Pulpillo, R. A., Goriely, S. & Janka, H.-T. Comprehensive nucleosynthesis analysis for ejecta of compact binary mergers. *Mon. Not. R. Astron. Soc*. **448**, 541–567 (2015).
35. Read, J. S. et al. Measuring the neutron star equation of state with gravitational wave observations. *Phys. Rev. D* **79**, 124033 (2009).
36. Lackey, B. D. & Wade, L. Reconstructing the neutron-star equation of state with gravitational-wave detectors from a realistic population of inspiralling binary neutron stars. *Phys. Rev. D* **91**, 043002 (2015).
37. Agathos, M. et al. Constraining the neutron star equation of state with gravitational wave signals from coalescing binary neutron stars. *Phys. Rev. D* **92**, 023012 (2015).
38. Abbott, B. P. et al. GW170817: measurements of neutron star radii and equation of state. *Phys. Rev. Lett*. **121**, 161101 (2018).
39. The LIGO Scientific Collaboration et al. A gravitational-wave standard siren measurement of the Hubble constant. *Nature* **551**, 85–88 (2017).
40. Hotokezaka, K. et al. A Hubble constant measurement from superluminal motion of the jet in GW170817. *Nat. Astron*. **3**, 940–944 (2019).
41. Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. Preprint at https://arxiv.org/abs/1807.06209 (2018).
42. Riess, A. G., Casertano, S., Yuan, W., Macri, L. M. & Scolnic, D. Large Magellanic Cloud Cepheid standards provide a 1% foundation for the determination of the Hubble constant and stronger evidence for physics beyond CDM. *Astrophys. J*. **876**, 85 (2019).
43. Damour, T. & Deruelle, N. General relativistic celestial mechanics of binary systems. I. The post-Newtonian motion. *Ann. I.H.P. Phys. Theor*. **43**, 107–132 (1985).
44. Damour, T. & Deruelle, N. General relativistic celestial mechanics of binary systems. II. The post-Newtonian timing formula. *Ann. I.H.P. Phys. Theor*. **44**, 263–292 (1986).
45. Damour, T. & Taylor, J. H. Strong-field tests of relativistic gravity and binary pulsars. *Phys. Rev. D* **45**, 1840–1868 (1992).
46. Yao, J. M., Manchester, R. N. & Wang, N. A new electron-density model for estimation of pulsar and FRB distances. *Astrophys. J*. **835**, 29 (2017).
47. Lorimer, D. R. & Kramer, M. *Handbook of Pulsar Astronomy* Vol. 4 (Cambridge Univ. Press, 2004).

Methods

Timing analysis

PSR J1913+1102 was discovered in 2012 by Einstein@Home in data from the PALFA survey[7], which uses the William E. Gordon 305-m radio telescope at the Arecibo Observatory in Puerto Rico to search for pulsars within 5° of the Galactic plane. Since the discovery, we have been using the Arecibo Observatory to regularly monitor this pulsar with the Mock Spectrometer and the PUPPI pulsar backend system. This has been done as both a dedicated follow-up campaign for PSR J1913+1102 and regularly as a test source before PALFA observing sessions.

The times of arrival of these pulses were measured by cross-correlating each pulse profile with a noise-free representative template profile, from which we calculated a phase shift and apply it to the observed time stamp of the data profile[48]. The uncertainty in each resulting pulse arrival time was determined by adopting the error in the calculated phase shift from the aforementioned correlation procedure. This process was carried out using the PSRCHIVE suite of analysis tools[49]. Corrections between terrestrial time and the observatory clock were applied using data from Global Positioning System satellites and the Bureau International des Poids et Mesures. Our model also included input from the Jet Propulsion Laboratory DE436 Solar System ephemeris, in order to convert measured arrival times to the reference frame of the Solar System barycentre, by taking into account the motion of Earth.

These barycentred times of arrival were then compared to a predictive model of their expected arrival at Earth using the TEMPO pulsar timing software package (https://github.com/nanograv/tempo). Every rotation of the neutron star was enumerated relative to a reference observing epoch by accounting in our model for intrinsic pulsar properties, such as the rotation frequency and its spin-down rate, as well as its sky position and proper motion. We also addressed potential arrival-time delays due to the frequency-dependent refractive effect of the ionized interstellar medium by including the dispersion measure (DM) in our timing model, which is the integrated column density of free electrons along the line of sight between Earth and the pulsar. The relatively large value DM = 339.026 ± 0.005 pc cm^{-3} for this pulsar explains why initial observations, primarily taken at an observing frequency centred at 1.4 GHz, displayed evidence of interstellar scattering that resulted in considerable smearing of the observed pulse shape due to multi-path propagation of the signal on its way to Earth[7]. This led to increased systematic uncertainties in the derived pulsar parameters, and we therefore switched to observations with the higher-frequency S-band Low receiver (centred at 2.4 GHz with a bandwidth of 800 MHz) to reduce these effects.

Along with these model parameters, our timing data resulted in an important measurement of the Keplerian orbital elements of the PSR J1913+1102 system, as well as several post-Keplerian parameters (these are quoted with 68% confidence-level uncertainties in Table 1). The latter are a theory-independent set of parameters that characterize perturbations on the Keplerian description of the orbit in the relativistic regime[43–45]. As described in the main text, we have now improved the measurement precision of the orbital precession rate and determined the Einstein delay. From this, we were able to constrain the individual masses of the neutron stars in this system by assuming that general relativity is the correct theory of gravity, as encapsulated in the DDGR timing model[43,44]. We have also made a precise determination of the orbital decay rate due to the emission of gravitational waves, which serve to remove orbital energy from the system over time. We expect the precision of the orbital decay to improve rapidly over time t (scaling as $t^{5/2}$) with further observations. It should also be noted that sources of kinematic biases can be introduced into the measured orbital decay and pulsar spin-down rates from apparent acceleration of the pulsar due to its tangential motion (that is, the Shklovskii effect[50]) and the Galactic potential[51,52]. We find the total proper motion of this pulsar to be 9.3 ± 0.9 mas yr^{-1}, within 3σ of what would be expected if the PSR J1913+1102 system were in the local standard of rest (6.50 mas yr^{-1}); this assumes a distance of 7.14 kpc. This distance is estimated from the measured value of DM, using a model of the Galactic free-electron density distribution[46]. The total kinematic bias to the observed orbital decay corresponds to approximately one-third of the uncertainty in the orbital decay measurement. We are therefore confident that our measurements are consistent with intrinsic parameter values for the pulsar at the current level of uncertainty.

Once we apply our model to the dataset, we produce post-fit timing residuals—the difference between the predicted and observed pulse arrival times (Fig. 3). The timing precision achieved by this fit to our pulse arrival time data is characterized by the root-mean-square (r.m.s.) of the post-fit timing residuals. Our analysis of the PSR J1913+1102 dataset resulted in r.m.s. residuals of 56.1 μs, consistent with the typical measured uncertainty in the observed pulse arrival times. We achieved a reduced χ^2 (that is, χ^2 divided by the number of degrees of freedom) of 1.01 for our fit, reaffirming the success of our timing model in describing the system, and implying that the timing residuals can be well represented by white Gaussian noise, as can be seen in Fig. 3.

Population synthesis

Modelling of the merger event that caused GW170817 has mostly relied on a DNS population consisting of roughly equal-mass neutron stars. Although this may be the result of a binary system with pre-merger mass ratio $q \approx 1$, the discovery of PSR J1913+1102 highlights the need to consider the effects of an asymmetric DNS merger. We note here for completeness that it is possible that GW170817 was caused by a neutron star–black hole merger. However, the abnormally low mass of the black hole in such a progenitor system would make this an unlikely scenario[53].

Previous studies[54–57] simulated the population of DNS binaries from the measured parameters of known systems within a modelled Galactic pulsar population; this was done using the known sensitivities of the pulsar survey in which they were discovered. The modelling must account for selection effects, including the search degradation factor due to orbital acceleration, calculated from a semi-analytical model with the pulsar and companion masses and the system inclination as input[58]. We calculated the probability density of the population of PSR J1913+1102-like DNSs that are beamed towards Earth ($N_{\mathrm{obs,J1913}}$) using the more precisely measured orbital properties presented in this work. Assuming a beaming correction fraction for the pulsar[56] of $f_{\mathrm{b}} = 4.6$, we derived the probability density of the total population ($N_{\mathrm{tot,J1913}}$) of J1913+1102-like DNS systems in the Galaxy: ($N_{\mathrm{pop,J1913}} = N_{\mathrm{obs,J1913}} \times f_{\mathrm{b}}$). The mode of the resulting distribution is $N_{\mathrm{tot,J1913}} = 700^{+2,600}_{-400}$, where the uncertainties represent the 90% confidence interval of the distribution. This is consistent with previous estimates[56], but has smaller error bars owing to the updated orbital parameters and the addition of a new radio pulsar survey[57].

Owing to its small orbital period, the PSR J1913+1102 system will merge in 470 Myr, well within one Hubble time. There are eight other known DNS systems in the Galaxy that will also merge within the age of the Universe (which we henceforth refer to as merging DNSs, MDNSs). We obtained the individual probability densities of the population of these MDNSs using results given in previous studies[56,57]. Assuming that these individual population distributions represent independent continuous random variables, we estimate the total population of MDNS systems in the Galaxy by convolving the individual population probability distributions, resulting in a mode $N_{\mathrm{tot,MDNS}} = (11.4^{+6.3}_{-3.8}) \times 10^3$. Using our derived probability densities of PSR J1913+1102-like systems together with those of all MDNS systems, we then compute the probability density of PSR J1913+1102-like DNS systems in the Galaxy as a fraction of the MDNS population to be 11^{+21}_{-9} % (90% confidence), using the median as the quoted value (with the mode of the distribution occurring at 6%); Fig. 2 presents the corresponding probability distribution. This in turn leads to an estimate that roughly one-tenth of detected DNS mergers result from the coalescence of an asymmetric binary system.

Data availability

All data are available from the corresponding author on request.

Code availability

The code used in this analysis is publicly available at: https://github.com/nanograv/tempo (pulsar timing analysis); https://github.com/NihanPol/2018-DNS-merger-rate (population synthesis); https://github.com/rferdman/pypsr (plotting tools).

48. Taylor, J. H. Pulsar timing and relativistic gravity. *Phil. Trans. R. Soc. A* **341**, 117–134 (1992).
49. Hotan, A. W., van Straten, W. & Manchester, R. N. PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis. *Publ. Astron. Soc. Pacif*. **21**, 302–309 (2004).
50. Shklovskii, I. S. Possible causes of the secular increase in pulsar periods. *Sov. Astron*. **13**, 562 (1970).
51. Nice, D. J. & Taylor, J. H. PSR J2019+2425 and PSR J2322+2057 and the proper motions of millisecond pulsars. *Astrophys. J*. **441**, 429–435 (1995).
52. Perera, B. B. P. et al. The dynamics of Galactic centre pulsars: constraining pulsar distances and intrinsic spin-down. *Mon. Not. R. Astron. Soc*. **487**, 1025–1039 (2019).
53. Kruckow, M. U., Tauris, T. M., Langer, N., Kramer, M. & Izzard, R. G. Progenitors of gravitational wave mergers: binary evolution with the stellar grid-based code COMBINE. *Mon. Not. R. Astron. Soc*. **481**, 1908–1949 (2018).
54. Kim, C., Kalogera, V. & Lorimer, D. R. The probability distribution of binary pulsar coalescence rates. I. Double neutron star systems in the Galactic field. *Astrophys. J*. **584**, 985–995 (2003).
55. Kim, C., Perera, B. B. P. & McLaughlin, M. A. Implications of PSR J0737-3039B for the Galactic NS-NS binary merger rate. *Mon. Not. R. Astron. Soc*. **448**, 928–938 (2015).
56. Pol, N., McLaughlin, M. & Lorimer, D. R. Future prospects for ground-based gravitational-wave detectors: the Galactic double neutron star merger rate revisited. *Astrophys. J*. **870**, 71 (2019); erratum 874, 186 (2019).
57. Pol, N., McLaughlin, M. & Lorimer, D. R. An updated Galactic double neutron star merger rate based on radio pulsar populations. *Res. Notes AAS* **4**, 22 (2020).
58. Bagchi, M., Lorimer, D. R. & Wolfe, S. On the detectability of eccentric binary pulsars. *Mon. Not. R. Astron. Soc*. **432**, 1303–1314 (2013).

Acknowledgements We thank S. Nissanke, T. Tauris and B. Metzger for constructive discussions, as well as C. Belczynski and T. Tauris for useful comments. The Arecibo Observatory is operated by the University of Central Florida, Ana G. Mèndez-Universidad Metropolitana and Yang Enterprises under a cooperative agreement with the National Science Foundation (NSF; AST-1744119). R.D.F. and P.C.C.F. acknowledge the support of the PHAROS COST Action (CA16214). S.C., J.M.C., F. Crawford, M.A.M. and N.P. are members of the NANOGrav Physics Frontiers Center, which is supported by NSF award number PHY-1430284. S.C. and J.M.C. also acknowledge support from NSF award AAG-1815242. J.W.T.H. acknowledges funding from an NWO Vici grant ('AstroFlash'). V.M.K. acknowledges support from an NSERC Discovery Grant and Herzberg Award, the Canada Research Chairs programme, the Canadian Institute for Advanced Research and FRQ-NT. E.P. is a Vanier Canada Graduate Scholar. I.H.S. acknowledges support for pulsar research at the University of British Columbia by an NSERC Discovery Grant and by the Canadian Institute for Advanced Research. J.v.L. acknowledges funding from an NWO Vici grant ('ARGO').

Author contributions R.D.F. contributed to timing analysis, wrote and ran mass probability contour code, led proposals for observing campaigns for this pulsar with the Arecibo telescope, composed the manuscript and produced Figs. 1, 3. P.C.C.F. led the timing analysis. B.P.P.P. and N.P. led the population synthesis analysis and produced Fig. 2. R.D.F., P.C.C.F., B.P.P.P., F. Camilo, S.C., J.M.C., F. Crawford, J.W.T.H., V.M.K., M.A.M., E.P., I.H.S. and J.v.L. contributed to the PALFA survey operations, as well as the discovery and follow-up observations of this pulsar. All authors contributed to discussion regarding the content of this manuscript.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.D.F.

Peer review information *Nature* thanks Chris Belczynski and Thomas Tauris for their contribution to the peer review of this work.

Tunable correlated states and spin-polarized phases in twisted bilayer–bilayer graphene

https://doi.org/10.1038/s41586-020-2260-6

Received: 24 March 2019

Accepted: 12 February 2020

Yuan Cao[1]✉, Daniel Rodan-Legrain[1], Oriol Rubies-Bigorda[1], Jeong Min Park[1], Kenji Watanabe[2], Takashi Taniguchi[2] & Pablo Jarillo-Herrero[1]✉

The recent discovery of correlated insulator states and superconductivity in magic-angle twisted bilayer graphene[1,2] has enabled the experimental investigation of electronic correlations in tunable flat-band systems realized in twisted van der Waals heterostructures[3–6]. This novel twist angle degree of freedom and control should be generalizable to other two-dimensional systems, which may exhibit similar correlated physics behaviour, and could enable techniques to tune and control the strength of electron–electron interactions. Here we report a highly tunable correlated system based on small-angle twisted bilayer–bilayer graphene (TBBG), consisting of two rotated sheets of Bernal-stacked bilayer graphene. We find that TBBG exhibits a rich phase diagram, with tunable correlated insulator states that are highly sensitive to both the twist angle and the application of an electric displacement field, the latter reflecting the inherent polarizability of Bernal-stacked bilayer graphene[7,8]. The correlated insulator states can be switched on and off by the displacement field at all integer electron fillings of the moiré unit cell. The response of these correlated states to magnetic fields suggests evidence of spin-polarized ground states, in stark contrast to magic-angle twisted bilayer graphene. Furthermore, in the regime of lower twist angles, TBBG shows multiple sets of flat bands near charge neutrality, resulting in numerous correlated states corresponding to half-filling of each of these flat bands, all of which are tunable by the displacement field as well. Our results could enable the exploration of twist-angle- and electric-field-controlled correlated phases of matter in multi-flat-band twisted superlattices.

Electronic correlations play a fundamental role in condensed-matter systems where the bandwidth is comparable to or less than the Coulomb energy between electrons. These correlation effects often manifest themselves as intriguing quantum phases of matter, such as ferromagnetism, superconductivity, Mott insulators or fractional quantum Hall states. Understanding, predicting and characterizing these correlated phases is of great interest in modern condensed-matter physics research and pose challenges to both experimentalists and theorists. Recent studies of twisted graphene superlattices have provided us with an ideal tunable platform to investigate electronic correlations in two dimensions[1,2,9–11]. Tuning the twist angle of two-dimensional (2D) van der Waals heterostructures to realize novel electronic states, an emerging field referred to as ‘twistronics’, has enabled physicists to explore a variety of novel phenomena[12–16]. When two layers of graphene are twisted by a specific angle, the phase diagram in the system exhibits correlated insulator states with similarities to Mott insulator systems[1,17], as well as unconventional superconducting states upon charge doping[2,9,11,18]. These effects might be originating from the many-body interactions between the electrons, when the band structure becomes substantially narrow as the twist angle approaches the first magic angle $\theta = 1.1°$ (refs. [3–5]).

Here we extend the twistronics research on graphene superlattices to a novel system with electrical displacement field tunability—twisted bilayer–bilayer graphene (TBBG), which consists of two sheets of untwisted Bernal-stacked bilayer graphene stacked together at an angle θ, as illustrated in Fig. 1a. The band structure of bilayer graphene is highly sensitive to the applied perpendicular electric displacement field[7,19,20], and therefore provides us with an extra knob to control the relative strength of electronic correlations in the bands[17]. Similar to twisted bilayer graphene (TBG)[3–5], the band structure of TBBG is flattened near about 1.1° (Fig. 2e–g)[21]. For devices with a twist angle near this value, our experiments show that the correlated insulator behaviour at $n_s/2$, $n_s/4$ and $3n_s/4$ can be sensitively turned on and off by the displacement field, where n_s is the density corresponding to fully filling one spin- and valley-degenerate superlattice band[22,23]. From their response to magnetic fields, all of these correlated states probably have a spin-polarized nature, with the $n_s/2$ state having a g-factor of about 1.5 for parallel fields, close to the bare electron spin g-factor of 2. In contrast, devices with a smaller twist angle of 0.84° show multiple displacement-field-tunable correlated states at higher fillings, consistent with the presence of several sets of correlated flat bands in the electronic structure. The combination of twist angle, electric displacement field and magnetic field provides a rich arena to investigate novel correlated phenomena in the emerging field of twistronics.

---

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]National Institute for Materials Science, Tsukuba, Japan. ✉e-mail: caoyuan@mit.edu; pjarillo@mit.edu

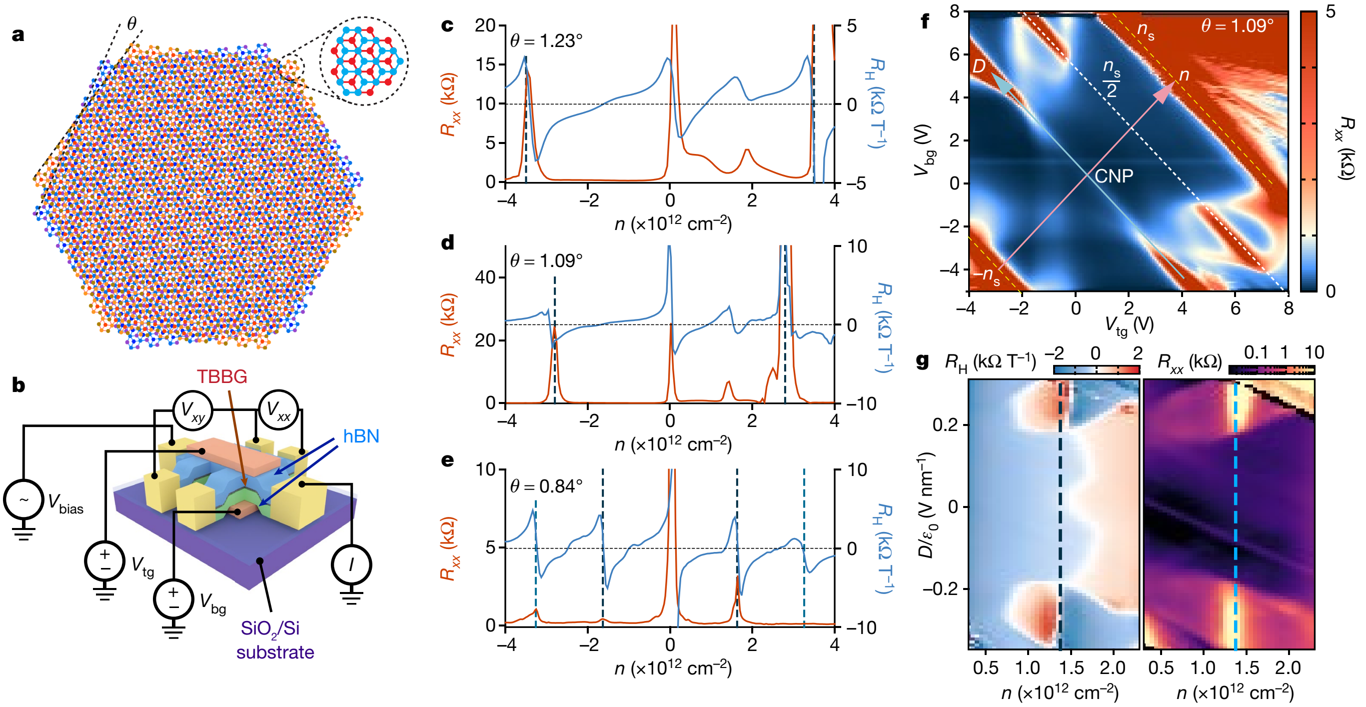


Fig. 1 | Structure and transport characterization of TBBG. a, TBBG consists of two sheets of Bernal-stacked bilayer graphene twisted at an angle θ. **b**, Schematic of a typical TBBG device with top and bottom gates and a Hall-bar geometry for transport measurements. **c**–**e**, Measured longitudinal resistance $R_{xx} = V_{xx}/I$ and low-field Hall coefficient $R_H = \mathrm{d}/\mathrm{d}B(V_{xy}/I)$ as functions of carrier density n in three devices with twist angles $\theta = 1.23°$ (**c**), 1.09° (**d**) and 0.84° (**e**). The vertical dashed lines denote multiples of the superlattice density n_s, where the peaking of R_{xx} and sign changing of R_H indicate the Fermi energy crosses a band edge of the superlattice bands. **f**, Resistance of the 1.09° TBBG device versus both V_{tg} and V_{bg}. The charge density n and displacement field D are related to the gate voltages by a linear transformation (Methods). The superlattice densities $\pm n_s$ and the half-filling at $n_s/2$ are indicated by dashed lines parallel to the D axis. Correlated insulator states are observed at $n_s/2$ filling in finite displacement fields. CNP, charge neutrality point. **g**, Map of low-field Hall coefficient R_H (left) and resistance R_{xx} (right) near the $n_s/2$ correlated states for the 1.09° TBBG device (the vertical dashed lines indicate $n_s/2$). We find that accompanying the onset of the correlated insulator states at $D/\varepsilon_0 \approx \pm 0.18$ V nm^{-1}, a new sign change of the Hall coefficient also emerges.

We fabricated high-mobility dual-gated TBBG devices with the previously reported 'tear and stack' method[22,23], using exfoliated Bernal-stacked bilayer graphene instead of monolayer graphene. The devices presumably have an AB–AB stacking configuration where the top and bottom bilayers retain the same AB stacking order, in contrast to the AB–BA structure that was predicted to show topological effects[24]. We measured the transport properties of six small-angle devices, and here we focus on three of the devices with twist angles $\theta = 1.23°$, 1.09° and 0.84° (see Extended Data Fig. 1 for other devices). The samples are all of high quality, as evident in the Landau fan diagrams, with Hall mobilities that can exceed 100,000 cm^2 V^{-1} s^{-1}, shown in Extended Data Fig. 2. Figure 1c–e shows the longitudinal resistance R_{xx} and the low-field Hall coefficient $R_H = \mathrm{d}R_{xy}/\mathrm{d}B$ versus charge density for these three devices at a temperature of $T = 4$ K, where B is the magnetic field perpendicular to the sample. In a superlattice, the electronic band structure is folded in the mini-Brillouin zone, defined by the moiré periodicity[4]. Each band in the mini-Brillouin zone can accommodate a total charge density of $n_s = 4/A$, where A is the size of the moiré unit cell and the pre-factor accounts for the spin and valley degeneracies[4,21,22]. The experimental results show a sign change in the Hall coefficient R_H at each multiple of n_s (vertical dashed lines in Fig. 1c–e), indicating the switching of hole-like pockets to electron-like pockets, and peaks in R_{xx}, indicating the crossing of new band edges (for $\theta = 0.84°$, the band edges at $-n_s$ and $\pm 2n_s$ may have only small gaps or may even be semi-metallic, and hence do not exhibit prominent peaks in R_{xx}). The sharpness of the peaks confirms that the devices exhibit relatively low disorder and have well-defined twist angles.

In the $\theta = 1.23°$ and $\theta = 1.09°$ devices, we observe signatures of newly formed gaps at $n_s/2$ when a displacement field D is applied perpendicular to the device. The dual-gate device geometry allows us to independently vary the total charge density n and D (see Methods for details of the transformation between gate voltages and (n, D)). Figure 1f shows the resistance map in the top gate voltage–bottom gate voltage (V_{tg}–V_{bg}) space for the $\theta = 1.09°$ device. At $D = 0$, no insulating behaviour other than the full-filling gaps at $\pm n_s$ is observed. However, when a displacement field D is applied in either direction, an insulating state appears at $n_s/2$ for a range of $|D|$. This new insulating state induced by the displacement field is further examined by measuring the Hall coefficient R_H versus n and D, as shown in the left panel of Fig. 1g ($\theta = 1.09°$ device), and comparing with R_{xx} shown in the right panel. At the onset of the insulating states at $D/\varepsilon_0 \approx \pm 0.18$ V nm^{-1}, where ε_0 is the vacuum permittivity, R_H develops additional sign changes adjacent to the insulating states, suggesting the creation of new gaps by the displacement field. The insulating states disappear when D/ε_0 exceeds ± 0.35 V nm^{-1}. In both the $\theta = 1.09°$ device and the $\theta = 1.23°$ devices, we find signatures of the onset of correlated behaviour at $n = -n_s/2$ and $D = 0$, but no well-developed insulating state is observed (Extended Data Fig. 1, Methods).

In the $\theta = 1.23°$ device, we observe a similar but more intricate hierarchy of tunable insulating states that stem from the interplay of correlations, the superlattice bands and the magnetic field. Figure 2a shows the n–D resistance map for the $\theta = 1.23°$ TBBG device measured at $T = 0.07$ K. Noticeably, as $|D|$ is increased, the insulating state at charge neutrality $n = 0$ strengthens in the same way as in the Bernal-stacked bilayer graphene[7,19,20], while the superlattice gaps at $\pm n_s$ weaken and eventually disappear (at $|D|/\varepsilon_0 > 0.6$ V nm^{-1} for the $+n_s$ insulating state and at $|D|/\varepsilon_0 > 0.35$ V nm^{-1} for the $-n_s$ insulating state). The band structures of TBBG in zero and finite external displacement fields calculated using

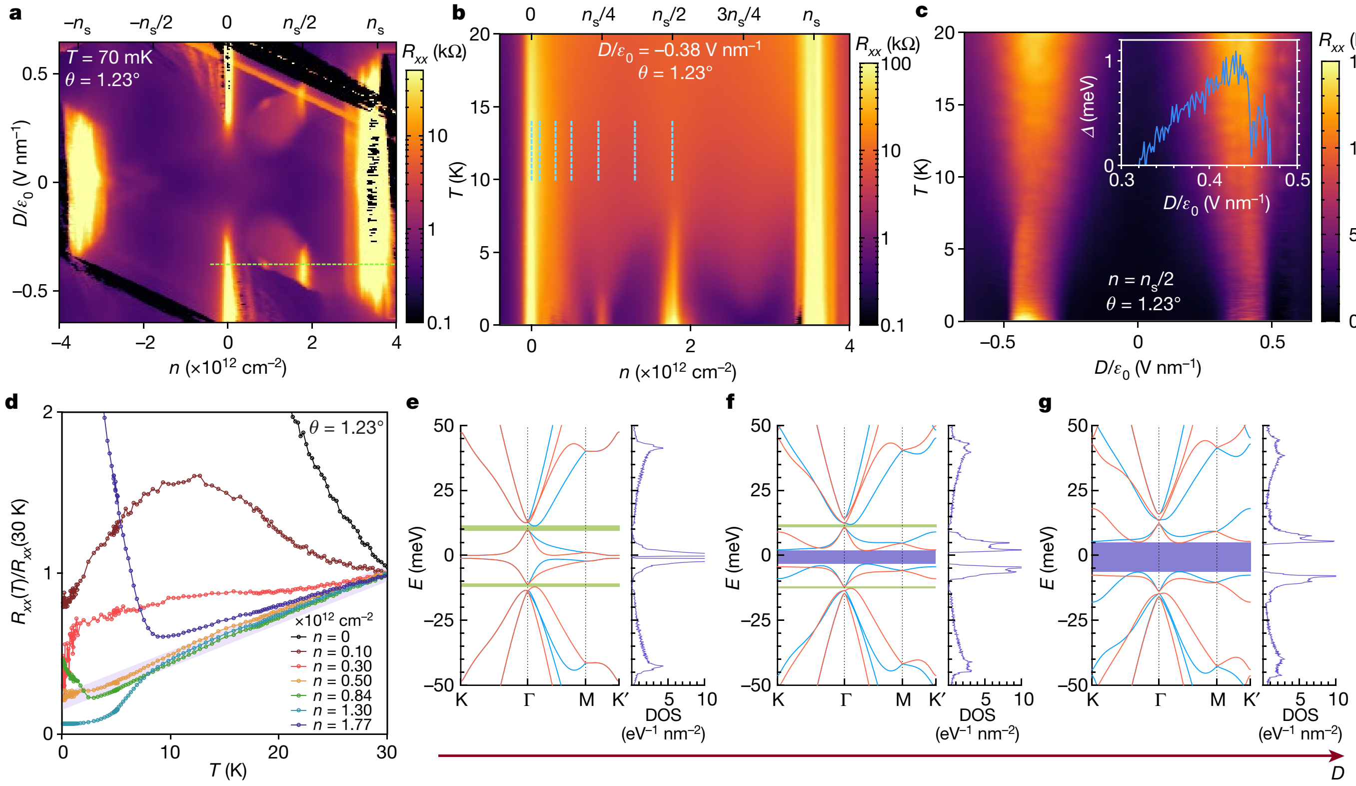


Fig. 2 | Displacement-field-tunable correlated insulator states in TBBG. **a**, Colour plot of resistance versus charge density n and displacement field D ($\theta = 1.23°$ device, section 1, see Methods). The green dashed line cutting through the $D < 0$ correlated state is the linecut along which **b** is taken (for the $\theta = 1.23°$ device, section 2, see Methods). **b**, Resistance versus n and T at a fixed $D/\varepsilon_0 = -0.38$ V nm^{-1}. The correlated insulator states at $n_s/4$ and $n_s/2$ are suppressed by increasing the temperature. **c**, Resistance at density $n_s/2$ versus displacement field and temperature. The resistance shows a maximum at approximately $D/\varepsilon_0 = \pm 0.4$ V nm^{-1}, the region where the correlated insulator state is present. The inset shows the thermal activation gap extracted from temperature dependence at different values of D across the $n_s/2$ state. **d**, Normalized resistance curves versus temperature at various densities between 0 and $n_s/2 \approx 1.77 \times 10^{12}$ cm^{-2}, which are indicated by dashed lines in **b**. Away from the charge neutrality point, all resistance curves show approximately linear R–T behaviour above 10 K, with similar slopes (Extended Data Fig. 3). **e**–**g**, Calculated band structure (left) and density of states (DOS; right) for $\theta = 1.23°$ TBBG at $\Delta V = 0$ (**e**), $\Delta V = 6$ mV (**f**) and $\Delta V = 12$ mV (**g**), where ΔV is the potential difference between adjacent graphene layers induced by the external displacement field (assumed to be the same between all layers). Single-particle bandgaps in the dispersion are highlighted green (below and above the flat bands) and purple (at charge neutrality) bars.

a continuum approximation are shown in Fig. 2e–g (see Methods for details). It should be noted that, although TBBG has twice the number of graphene layers than TBG, the band counting is the same, that is, each band (spin and valley degenerate) accommodates four electrons per moiré unit cell. At zero displacement field, the calculated gap at the charge neutrality is negligible, while the superlattice gaps above and below the flat bands are non-zero. When the displacement field is increased, the charge neutrality gap quickly widens while the superlattice gaps become smaller and eventually vanish, in agreement with our experimental observations.

At intermediate displacement fields around $D/\varepsilon_0 = -0.38$ V nm^{-1}, we observe the insulating states not only at $n_s/2$ over a wider range of D, but also at $n_s/4$ over a smaller range (Fig. 2a). We attribute these states to a Mott-like mechanism similar to those observed in TBG, which results from the Coulomb repulsion of the electrons in the flat bands when each unit cell hosts exactly one or two electrons, corresponding to $n_s/4$ and $n_s/2$ fillings, respectively. The $n_s/4$ state requires a finer tuning of D to be revealed, possibly due to the smaller gap size. This is evident from Fig. 2b, where we show the resistance versus n and temperature T with the displacement field D/ε_0 fixed at −0.38 V nm^{-1}. While the $n_s/2$ state persists up to approximately 8 K, the $n_s/4$ state disappears at less than 3 K, indicating a smaller gap. Figure 2c shows the resistance of the $n_s/2$ state versus the displacement field and temperature. The 'optimal' displacement field to reach the maximal resistance is approximately ±0.4 V nm^{-1}. As the temperature increases, the peak in R_{xx} not only decreases in value but also broadens in D. In the inset, we show the evolution of the gap versus the displacement field. At temperatures higher than 10 K and away from the charge neutrality point, the transport is dominated by a linear R–T behaviour similar to that observed in TBG (Fig. 2d, see also Extended Data Fig. 3, Methods)[2,9,25,26].

Figure 3 shows the response of the various correlated states to magnetic fields in the perpendicular or in-plane direction with respect to the sample plane. Figure 3a–c shows the n–D maps of the resistance for the $\theta = 1.23°$ device at $B = 0$ T, $B_\perp = 8$ T and $B_\parallel = 8$ T, respectively. The plots focus on densities from $n = 0$ to $n = n_s$. Figure 3a shows the band insulator states at $n = 0$ and $n = n_s$, as well as the correlated insulating states at $n_s/2$ and $n_s/4$ (encircled by dashed lines), but not at $3n_s/4$ filling at this zero magnetic field. Interestingly, at $B_\perp = 8$ T (Fig. 3b), the correlated insulating states at $n_s/4$ and $n_s/2$ vanish at their original positions centred around $D/\varepsilon_0 = -0.38$ V nm^{-1}, whereas new insulating states appear at $n = n_s/4$, $D/\varepsilon_0 \approx -0.2$ to -0.35 V nm^{-1}, and $n = n_s/2$, $D/\varepsilon_0 \approx -0.45$ to -0.6 V nm^{-1}, above and below their original positions at $B = 0$, respectively. A new correlated insulating state also now appears at $3n_s/4$, $D/\varepsilon_0 \approx -0.4$ to -0.5 V nm^{-1}. However, no such strong shift is observed with in-plane magnetic field (Fig. 3c). At $B_\parallel = 8$ T, the correlated insulating states are clearly visible at all integer electron fillings ($n_s/4$, $n_s/2$, $3n_s/4$) near $D/\varepsilon_0 = -0.38$ V nm^{-1}. Figure 3d, e shows the evolution of the $n_s/2$ insulating state as a function of $B_\perp$ and $B_\parallel$. An abrupt

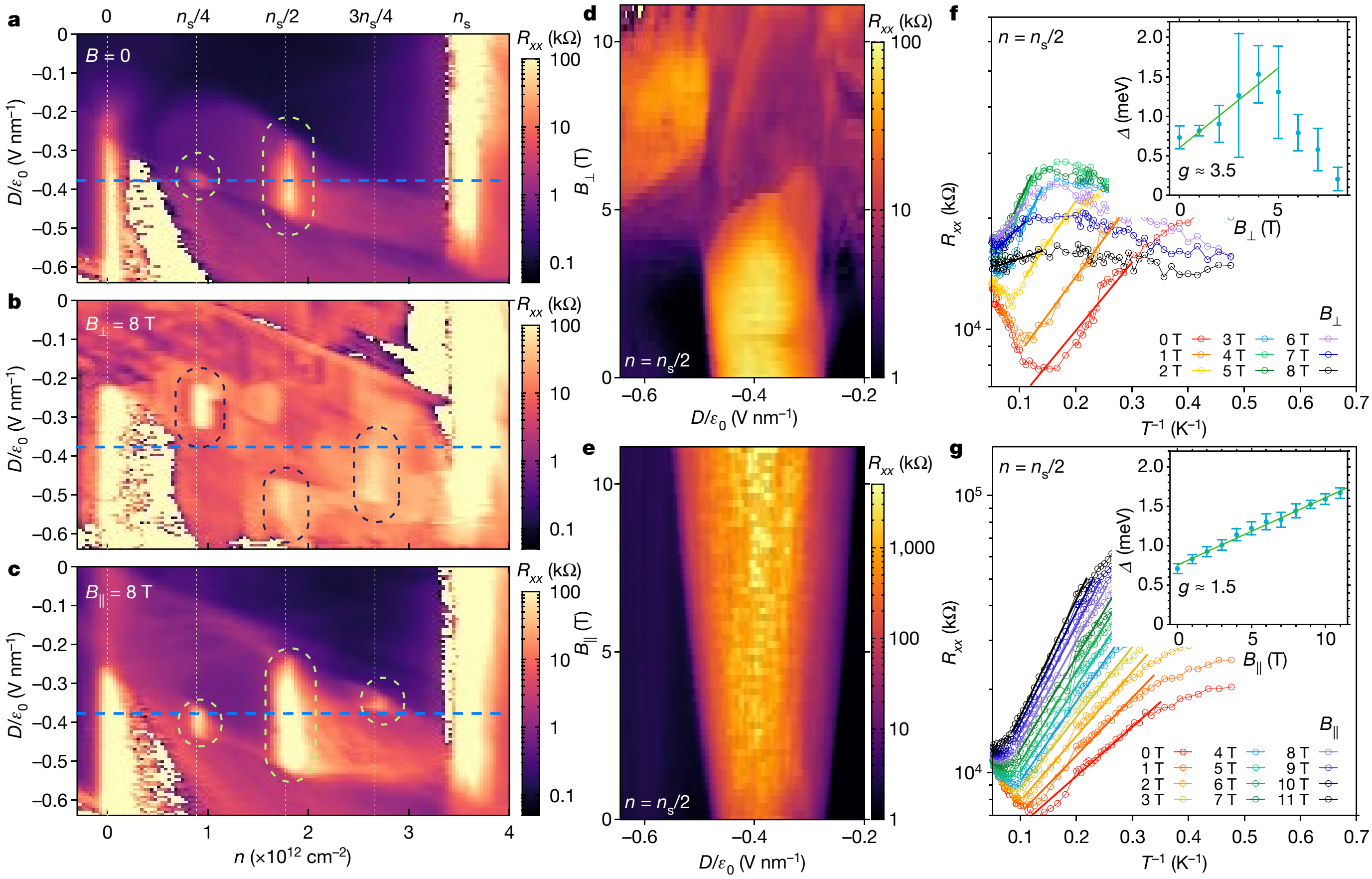


Fig. 3 | Magnetic field response of the displacement-field-tunable correlated insulator states in TBBG. a–c, Resistance plot for the $\theta = 1.23°$ TBBG device in magnetic fields of $B = 0$ (**a**), $B_\perp = 8$ T perpendicular to the sample (**b**) and $B_\parallel = 8$ T parallel to the sample (**c**). All measurements are taken at sample temperature $T = 0.07$ K. Various correlated states at integer electron fillings of the moiré unit cell are indicated by dashed circles. At zero field, only the $n_s/4$ and $n_s/2$ states appear around $|D|/\varepsilon_0 = 0.38$ V nm^{-1} (denoted by blue dashed lines). In a perpendicular field of 8 T, the $n_s/4$ state shifts towards lower $|D|$, the $n_s/2$ state shifts towards higher $|D|$ and a $3n_s/4$ state also emerges. In a parallel field of 8 T, however, the position of the states barely shifts but their resistance increases monotonically. **d**, **e**, Resistance at $n = n_s/2$ versus displacement field and magnetic field applied perpendicular (**d**) and in-plane (**e**) with respect to the device. While the correlated insulator state monotonically strengthens in $B_\parallel$, the perpendicular field induces a phase transition at around $B_\perp = 5$ T, where the correlated state abruptly shifts to higher $|D|$. **f**, **g**, Temperature dependence of the resistance at the $n_s/2$ insulator in perpendicular (**f**) and in-plane (**g**) magnetic fields. The insets show the thermal activation gaps extracted from the Arrhenius fits ($R \approx e^{-\frac{\Delta}{2k_BT}}$, where k_B is the Boltzmann constant) in the main figures (solid lines) versus the magnitude of the field in the respective orientation. Error bars correspond to a confidence level of 0.99. The linear fit of the thermal activation gap gives a g-factor of about 3.5 for the perpendicular field (up to 5 T only) and 1.5 for the in-plane field (entire field range).

shift in the range of D for which the insulating state appears occurs at $B_\perp = 5$ T, whereas the insulating state strengthens monotonically with the in-plane magnetic field.

The key difference between the effects of the perpendicular and in-plane magnetic fields lies in the fact that the lateral dimension of the unit cell in TBBG, about 10 nm, is much larger than the thickness of the system, about 1 nm. Therefore, while both fields couple equally to the spins of the correlated electrons, $B_\parallel$ has a much weaker (but non-zero) effect on the orbital movement of the electrons. To theoretically understand the behaviour of the correlated insulating states in a magnetic field, we first have to identify their ground state. Figure 3f, g shows the evolution of the thermal activation gap of the $n_s/2$ state in both $B_\perp$ and $B_\parallel$. We find a g-factor of $g_\perp \approx 3.5$ for the perpendicular direction (up to 5 T before the shift occurs) and a g-factor of $g_\parallel \approx 1.5$ for the in-plane direction. $g_\parallel$ is close to (but less than) $g = 2$, which is expected for a spin-polarized ground state with contribution from only the electron spins. This difference is theoretically expected because of finite in-plane orbital effects[27]. Therefore, on the basis of these measurements, we may conclude that the correlated insulating states have a spin-polarized nature. These observations establish TBBG as a distinctive system from the previously reported magic-angle TBG system[1,2,9], which exhibits half-filling insulating states that are shown to be spin unpolarized, as they are suppressed by an in-plane magnetic field. In $B_\perp$, however, one would expect orbital effects to have a more substantial role. We may attribute the larger $g_\perp$ of about 3.5 to exchange-induced enhancement effects, similar to what is observed in Landau levels of gallium arsenide quantum wells and graphene[28,29]. In Extended Data Fig. 4, we provide additional magnetic field response data for the $n_s/4$ and the $3n_s/4$ states. Both of these states also exhibit a spin-polarized behaviour, as they become more resistive under the in-plane magnetic field.

In addition to the discussion above, we noticed that all the correlated insulating states in the $\theta = 1.23°$ TBBG device, whether at zero magnetic field or high magnetic fields, lie within the range $D/\varepsilon_0 \approx -0.6$ to -0.2 V nm^{-1}. Coincidentally, this is also the range where both the gap at the charge neutrality ($n = 0$) and the gap at the superlattice density ($n = n_s$) are well developed (that is, the case in Fig. 2f). On the basis of this observation, we suggest that the displacement field tunability of the correlated states is tied to the modulation of the single-particle bandgaps by the displacement field[27]. When either gap at $n = 0$ or $n = n_s$ is absent, the thermally excited or disorder-scattered carriers from the upper or lower bands would suppress the ordering of the electrons and hence the correlated states. Further theoretical

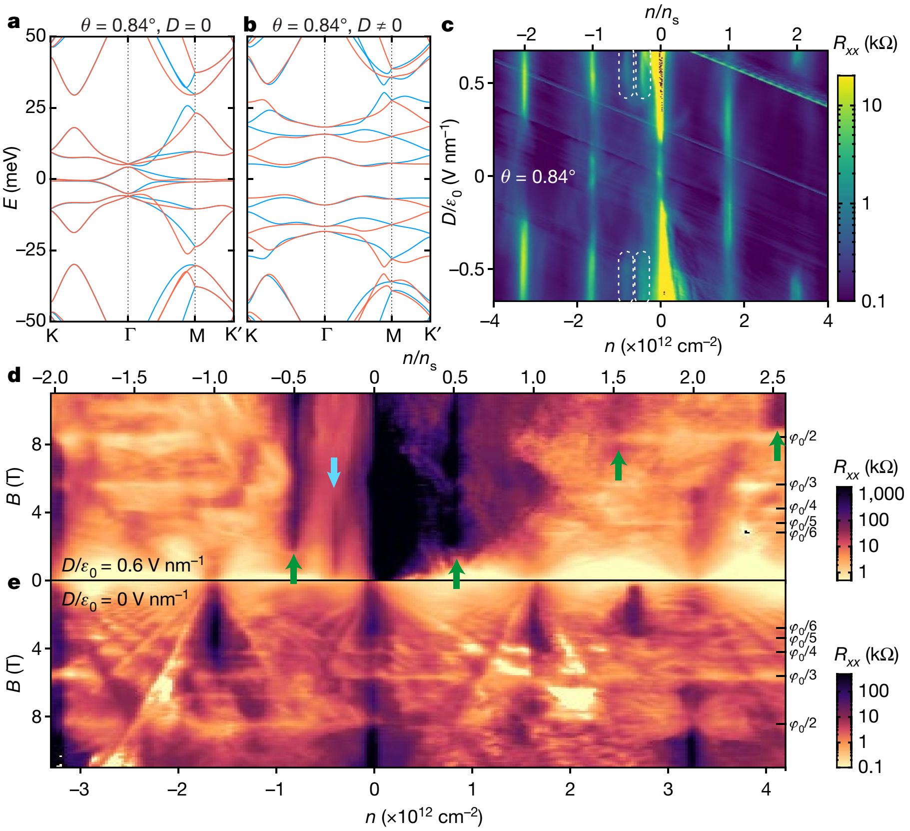


Fig. 4 | Correlated insulator states in a multi-flat-band system. **a**, **b**, Calculated band structure of $\theta = 0.84°$ TBBG without an interlayer potential (**a**) and with an interlayer potential $\Delta V = 18$ mV (**b**). Near charge neutrality, within a 50-meV window, there are in total six sets of flat bands spanning densities $-3n_s$ to $3n_s$. Upon applying a displacement field, these bands are further flattened and separated from each other, which makes them more prone to giving rise to correlated states at each half-filling. **c**, Resistance map of a $\theta = 0.84°$ TBBG device measured at $T = 0.07$ K. The top axis is the charge density normalized to the superlattice density n_s. Besides the D-tunable gaps at multiples of n_s, we find signatures of correlated states at $n/n_s = -1/2, -1/4$ for $|D|/\varepsilon_0 > 0.4$ V nm^{-1}, which are indicated by dashed circles. **d**, Resistance as a function of charge density and perpendicular magnetic field B when a displacement is present, $D/\varepsilon_0 = 0.6$ V nm^{-1}. We find clear correlated states at $n/n_s = -1/2, 1/4$ and $1/2$, and also evidences at 3/2 and 5/2 fillings, as indicated by arrows (blue and green arrows indicate half-fillings and quarter-filling, respectively). **e**, For comparison, when no displacement field is present, we do not find any signature of half-filling correlated states. Owing to the formation of a superlattice, we also observe Hofstadter butterfly related features when B is such that the magnetic flux in each unit cell is equal to $\varphi_0/2, \varphi_0/3, \varphi_0/4$ and so on, where $\varphi_0 = h/e$ is the flux quantum, h and e being Planck's constant and electron charge, respectively.

work is needed to reveal the detailed structure of the displacement field dependence of the correlated states.

We have also investigated the regime of substantially smaller twist angles. Unlike the case of TBG, further reduction of the twist angle of TBBG to 0.84° results not in one, but rather three pairs of flat bands, separated from other bands by bandgaps (Fig. 4a). The application of an electrical displacement field further flattens these bands and separates them from each other (Fig. 4b). This would imply that all electrons within the density range $-3n_s$ to $+3n_s$ might experience strong Coulomb interactions and that their correlations can get further enhanced by applying a displacement field. These predictions from the band theory are consistent with our experimental observations. In Fig. 4c, where we show the resistance map of the $\theta = 0.84°$ TBBG device versus n and D, we indeed find that the weak signatures of the $-n_s/2$ and $-n_s/4$ correlated insulating states appear only at high displacement fields $|D|/\varepsilon_0 > 0.4$ V m^{-1} (encircled by white dashed lines). The full-filling gaps at $\pm n_s$ and $\pm 2n_s$ are tunable by the displacement field to different extents as well.

As we turn on a perpendicular magnetic field, a series of correlated insulator states appear across the entire density range spanning the multiple flat bands. Figure 4d, e shows the Landau fan diagrams at $D/\varepsilon_0 = 0.6$ V m^{-1} and $D = 0$, respectively. At zero displacement field, the Landau fan shows a complicated Hofstadter butterfly pattern due to commensurate flux threading into the unit cell[12–14] (see also Methods, Extended Data Fig. 2), but no correlated state is observed at half-fillings or quarter-fillings. We note that a resistive region appears at $n \approx 1.63n_s$ in Fig. 4e, which does not coincide with any commensurate filling and might be ascribed to twist-angle inhomogeneity in the sample. In contrast, at $D/\varepsilon_0 = 0.6$ V nm^{-1}, we find clear signatures of correlated states at $n_s/2$ and $-n_s/2$ in the centre flat bands, and weak evidences at $3n_s/2$ and $5n_s/2$ in the upper flat bands. All of these half-filling correlated states appear to be enhanced by the application of a perpendicular magnetic field, which we attribute to the same spin/orbital combined enhancement of the correlated gaps as in the $n_s/2$ state of the $\theta = 1.23°$ device (Fig. 3f). The correlated states at $\pm n_s/2$ appear to be much stronger than the states at $3n_s/2$ and $5n_s/2$ in high magnetic fields, consistent with the fact that from our calculations, the pair of bands closer to charge neutrality is much flatter than the other two pairs farther away from charge neutrality, as can be seen in Fig. 4b. The resistance of the quarter-filling state at $n_s/4$, however, does not increase monotonically with the perpendicular field, but rather eventually gets suppressed at 5 T.

Our results show that TBBG exhibits a rich spectrum of correlated phases tunable by twist angle, electric displacement field and magnetic field, enabling further studies of strongly correlated physics and topology in multi-flat-band systems[21].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2260-6.

1. Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
2. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
3. Suárez Morell, E., Correa, J. D., Vargas, P., Pacheco, M. & Barticevic, Z. Flat bands in slightly twisted bilayer graphene: tight-binding calculations. *Phys. Rev. B* **82**, 121407 (2010).
4. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
5. Lopes dos Santos, J. M. B., Peres, N. M. R. & Castro Neto, A. H. Continuum model of the twisted graphene bilayer. *Phys. Rev. B* **86**, 155449 (2012).
6. Carr, S. et al. Twistronics: manipulating the electronic properties of two-dimensional layered structures through their twist angle. *Phys. Rev. B* **95**, 075420 (2017).
7. McCann, E. & Koshino, M. The electronic properties of bilayer graphene. *Rep. Prog. Phys.* **76**, 056503 (2013).
8. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009).
9. Yankowitz, M. et al. Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).
10. Sharpe, A. L. et al. Emergent ferromagnetism near three-quarters filling in twisted bilayer graphene. *Science* **365**, 605–608 (2019).
11. Lu, X. et al. Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene. *Nature* **574**, 653–657 (2019).
12. Hunt, B. et al. Massive Dirac fermions and Hofstadter butterfly in a van der Waals heterostructure. *Science* **340**, 1427–1430 (2013).
13. Ponomarenko, L. A. et al. Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
14. Dean, C. R. et al. Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
15. Krishna Kumar, R. et al. High-temperature quantum oscillations caused by recurring Bloch states in graphene superlattices. *Science* **357**, 181–184 (2017).
16. Li, G. et al. Observation of Van Hove singularities in twisted graphene layers. *Nat. Phys.* **6**, 109–113 (2010).
17. Chen, G. et al. Evidence of a gate-tunable Mott insulator in a trilayer graphene moiré superlattice. *Nat. Phys.* **15**, 237–241 (2019).
18. Chen, G. et al. Signatures of tunable superconductivity in a trilayer graphene moiré superlattice. *Nature* **572**, 215–219 (2019).
19. Oostinga, J. B., Heersche, H. B., Liu, X., Morpurgo, A. F. & Vandersypen, L. M. K. Gate-induced insulating state in bilayer graphene devices. *Nat. Mater.* **7**, 151–157 (2008).
20. Zhang, Y. et al. Direct observation of a widely tunable bandgap in bilayer graphene. *Nature* **459**, 820–823 (2009).
21. Zhang, Y.-H., Mao, D., Cao, Y., Jarillo-Herrero, P. & Senthil, T. Nearly flat Chern bands in moiré superlattices. *Phys. Rev. B* **99**, 075127 (2019).
22. Cao, Y. et al. Superlattice-induced insulating states and valley-protected orbits in twisted bilayer graphene. *Phys. Rev. Lett.* **117**, 116804 (2016).
23. Kim, Y. et al. Charge inversion and topological phase transition at a twist angle induced Van Hove singularity of bilayer graphene. *Nano Lett.* **16**, 5053–5059 (2016).
24. Koshino, M. Band structure and topological property of twisted double bilayer graphenes. *Phys. Rev. B* **99**, 235406 (2019).
25. Polshyn, H. et al. Large linear-in-temperature resistivity in twisted bilayer graphene. *Nat. Phys.* **15**, 1011–1016 (2019).
26. Cao, Y. et al. Strange metal in magic-angle graphene with near Planckian dissipation. *Phys. Rev. Lett.* **124**, 076801 (2020).
27. Lee, J. Y. et al. Theory of correlated insulating behaviour and spin-triplet superconductivity in twisted double bilayer graphene. *Nat. Commun.* **10**, 5333 (2019).
28. Raymond, A. et al. Gigantic exchange enhancement of spin *g*-factor for two-dimensional electron gas in GaAs. *Solid State Commun.* **55**, 271–274 (1985).
29. Young, A. F. et al. Spin and valley quantum Hall ferromagnetism in graphene. *Nat. Phys.* **8**, 550–556 (2012).

Methods

Fabrication and measurement

The reported devices were fabricated with two sheets of Bernal-stacked bilayer graphene and encapsulated by two hexagonal boron nitride (hBN) flakes. Both bilayer graphene and hBN were exfoliated on SiO_2/Si substrates, and the thickness and quality of the flakes were confirmed with optical microscopy and atomic force microscopy. A modified polymer-based dry pick-up technique was used for the fabrication of the heterostructures. A poly(bisphenol A carbonate) (PC)/polydimethylsiloxane (PDMS) layer on a glass slide was positioned in the micro-positioning stage to first pick up an hBN flake at around 100 °C. The van der Waals interaction between the hBN and bilayer graphene then allowed us to tear the bilayer graphene flake, which was then rotated at a desired angle and stacked at room temperature. The resulting hBN/bilayer graphene/bilayer graphene heterostructure was released on another hBN flake on a palladium/gold back gate that was pre-heated to 170 °C, using a hot-transfer method[30,31]. The desired geometry of the devices was achieved with electron beam lithography and reactive ion etching. The electrical contacts and top gate were deposited by thermal evaporation of chromium/gold, making edge contacts to the encapsulated graphene[32].

Electronic transport measurements were performed in a dilution refrigerator with a superconducting magnet, with a base electronic temperature of 70 mK. The data were obtained with low-frequency lock-in techniques. We measured the current through the sample amplified by 10^7 V A^{-1} and the four-probe voltage amplified by 1,000, using SR-830 lock-in amplifiers that were all synchronized to the same frequency between around 1 and 20 Hz. For resistance measurements, we typically used a voltage excitation of less than 100 µV or current excitation of less than 10 nA.

List of measured TBBG devices

Following the definition given in the main text and accounting for offsets in the gate voltages due to impurity doping, n and D are related to the top and bottom gate voltages V_{tg} and V_{bg} by

$$n = [c_{tg}(V_{tg} - V_{tg,0}) + c_{bg}(V_{bg} - V_{bg,0})]/e$$

$$D = [-c_{tg}(V_{tg} - V_{tg,0}) + c_{bg}(V_{bg} - V_{bg,0})]/2$$

Extended Data Table 1 lists the twist angles and parameters c_{tg} (top gate capacitance per area), c_{bg} (bottom gate capacitance per area), $V_{tg,0}$ (top gate voltage offset), $V_{bg,0}$ (bottom gate voltage offset) and n_s (superlattice density) for all devices discussed in this work, including those shown in the Extended Data figures. e is unit electron charge. These parameters are estimated to satisfy that all diagonal features in the V_{tg}–V_{bg} maps are rotated to be vertical in the corresponding n–D maps, and the features should be symmetrical with respect to $D = 0$ after the transformation.

In Extended Data Fig. 1a–f, we show V_{tg}–V_{bg} resistance maps for all six TBBG devices we measured. Extended Data Fig. 1c, d was measured in the same TBBG sample, but in different sample regions that are approximately 27 µm apart (sections 1 and 2, respectively). Both regions have identical parameters (hence the two identical rows in Extended Data Table 1), with the same twist angle $\theta = 1.23°$, and also nearly identical transport characteristics. The two sections are electrically disconnected via etching, but the extracted twist angles from the data have a difference of less than 0.01°, suggesting very uniform twist angles across this entire sample.

In almost all TBBG samples, we noticed a peculiar cross-like pattern around $(n, D) = (-n_s/2, 0)$, that is, near p-side half-filling of the superlattice band. This is especially apparent in the 1.09° and 1.23° devices, which are highlighted in Extended Data Fig. 1g, h. The p-side band does not exhibit a strong D-tunable correlated state as elaborated in the main text, possibly due to the larger bandwidth compared with its n-side counterpart. This cross-like pattern might represent an onset of correlated behaviour near half-filling of the band. Further experimental work and theoretical insight are needed to understand this phenomenon.

Sample quality and Landau fans

To demonstrate the high quality of our fabricated TBBG devices, we measured the Landau fan diagrams and Hall mobilities of all three devices discussed in the main text, as shown in Extended Data Fig. 2. The Hall mobilities are extracted from the ratio between the Hall coefficient R_H and longitudinal resistance at small magnetic fields ($B < 0.5$ T). All three samples exhibit high Hall mobilities close to or above 100,000 $cm^2 V^{-1} s^{-1}$.

All three devices also show clear Landau fans starting from about 1 T. The filling factor of each level is labelled in the lower panels of each plot. In particular, due to the lower angle of the $\theta = 0.84°$ device, its Landau fan displays a complicated Hofstadter's butterfly pattern starting from 3 T.

Linear *R*–*T* behaviour

Extended Data Fig. 3 shows the resistance versus temperature behaviour, at different densities, observed across several small-angle TBBG devices. In the 1.23° device, we find approximately linear R–T behaviour above 10 K for densities ranging from around 0.5×10^{12} to 2.5×10^{12} cm^{-2}, encompassing the $n_s/2$ correlated state. The resistance slope in this range of densities does not vary very substantially, ranging from around 210 to 350 Ω K^{-1}. As all our devices have length-to-width ratios close to one, these slope values are therefore close to those reported in TBG[25,26]. In stark contrast, the resistance behaviour in the hole-doping side ($n < 0$), as shown in Extended Data Fig. 3b, shows qualitatively different behaviour: it does not show linear R–T characteristics, at least up to 30 K, and the resistance value is about an order of magnitude smaller than on the electron-doping side. These data are consistent with the picture that the electron-doping band is flatter than the hole-doping band, therefore exhibiting more pronounced correlated phenomena, examples being the $n_s/2$ insulator state and the linear R–T behaviour. Extended Data Fig. 3c shows R–T curves close to the $n_s/2$ state.

The data for the 1.09° device show a similar trend of linear R–T behaviour starting around 5–10 K, as shown in Extended Data Fig. 3d.

In the 0.84° device, we find a very different behaviour. There is a region of sublinear or approximately linear R–T behaviour at all densities, except at multiples of n_s, but the resistance slope is now strongly dependent on the charge density n. The slope approximately follows a power law $\frac{dR_{xx}}{dT} \propto n^a$ where $a \approx -1.77$ (see inset).

Theoretical methods

The band structures shown in the main text are calculated using a continuum model based on the original continuum model for TBG[4,5], which qualitatively captures most of the important features of the bands in TBBG including displacement-field dependence. To the lowest order, the continuum model of twisted graphene superlattices is built on the approximation that the interlayer coupling between the A/B sublattice of one layer and the A/B sublattice of the other layer has a sinusoidal variation over the periodicity of the moiré pattern. For the three possible directions of interlayer connections between the wave vectors in the Brillouin zone, there are three connection matrices

$$H_1 = w\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$H_2 = w\begin{pmatrix} \omega^2 & 1 \\ \omega & \omega^2 \end{pmatrix}$$

$$H_3 = w\begin{pmatrix}\omega & 1\\ \omega^2 & \omega\end{pmatrix}$$

where w is the interlayer hopping energy and $\omega = \exp(2\pi i/3)$. $H_{i,\alpha\beta}$, with $\alpha\beta$ = A, B represents the hopping between sublattice α in the first layer to sublattice β in the second layer, with momentum transfer determined by i (see ref. [4] for definition). Note that in this gauge choice, the origin of rotation is chosen where the B sublattice of the first layer coincides with the A sublattice of the second layer, so that the $H_{i,\mathrm{BA}}$ component has zero phase while the other terms acquire phases. A different gauge choice is equivalent to an interlayer translation, which has been shown to have a negligible effect in the case of small twist angles[4,5].

To extend this formulation to TBBG, we add a simplified bilayer graphene Hamiltonian

$$H_\mathrm{b} = \begin{pmatrix}0 & 0\\ w_\mathrm{b} & 0\end{pmatrix}$$

between the non-twisted layers. The momentum transfer is zero since the bilayers are not twisted and the coupling is constant over the moiré unit cell. For simplicity, we consider only the 'dimer' coupling in the bilayer, neglecting second-nearest-neighbour hopping terms and trigonal warping terms. The two bilayers in TBBG (layers 1–2 and layers 3–4) have the same stacking order, that is, for zero twist angle the total stacking would be 'ABAB' instead of 'ABBA'. In the calculations used in the main text, we used parameters w = 0.1 eV and w_b = 0.4 eV, so that when either parameter is turned off we obtain either the two non-interacting bilayer graphene (w = 0) or the non-interacting TBG and two-monolayer graphene (w_b = 0).

Additional magnetic-field-response data

Extended Data Fig. 4 shows the response of correlated states at $n_s/4$ and $3n_s/4$ in a perpendicular or in-plane magnetic field, similar to Fig. 3d, e, for the $\theta = 1.23°$ device. For the $n_s/4$ state, we also find a signature of a phase transition at $D/\varepsilon_0 = -0.36\ \mathrm{V\ nm^{-1}}$, manifesting as a shift of the D location of the correlated insulator as $B_\perp$ exceeds 6 T. The $3n_s/4$ state shows an overall monotonic increase of resistance and exhibits no shift in the position in D. In an in-plane field, however, as shown in Extended Data Fig. 4b, d, both quarter-filling states show a monotonic enhancement as $B_\parallel$ is increased, suggesting that they may have a similar spin-polarized ground state as the $n_s/2$ state.

Current–voltage curves and the impact of excitation current on *g*-factor

In Extended Data Fig. 5, we have plotted the current–voltage (I–V) curves and differential resistance curves of the $\theta = 1.23°$ device when it is in the correlated insulator states at $n_s/4$ and $n_s/2$. In the insulator states, we find a highly nonlinear region near zero d.c. bias $I_\mathrm{b} = 0$ where the differential resistance $\mathrm{d}V_{xx}/\mathrm{d}I_\mathrm{b}$ is substantially enhanced. This is in agreement with the existence of a small energy gap, which is overcome at higher bias voltages/currents. Outside of the insulator regions (such as shown in Extended Data Fig. 5b), the I–V curves are mostly linear. For measuring the g-factors at $n_s/2$, we therefore used a much smaller excitation current of 0.1 nA to truthfully measure the differential resistance at $I_\mathrm{b} = 0$.

We comment here on the effect of the a.c. excitation current on the measured gap sizes and the g-factor. When sourcing an a.c. bias current to measure the resistance using a lock-in technique, we effectively measure a weighted average of the differential resistance near zero bias. Owing to the highly nonlinear I–V curve at the $n_s/2$ state, if the a.c. excitation is large, this average value will be much less than the peak value. Furthermore, the average value measured in this case can have a very different temperature dependence compared with the zero-bias value. For example, although to the best of our knowledge there is no detailed analysis of the high-bias behaviour in the correlated insulator state of TBG, TBBG or related systems, if one considers the high-bias transport to have a contribution from a mechanism similar to Zener breakdown in semiconductors in an electrical field, the current is essentially independent of the temperature. There could be other contributions to the high-bias transport as well, but in general their temperature dependence would not be identical to the zero-bias peak. In the Arrhenius fit that we use to extract the gap size, the gap size Δ is basically equal to how fast the resistance exponentially rises with T^{-1}. Therefore, a reduction of temperature dependence means that by averaging the higher bias differential resistance one would substantially underestimate the energy gap Δ, and also the g-factor $g \propto \delta\Delta/\delta B$.

In Extended Data Fig. 6, we compare the Arrhenius fits of the resistance at $n_s/2$ and $n_s/4$ states, using a small excitation (0.1 nA) and a larger excitation (around 5–10 nA). We indeed find that by using an excessive excitation, both the gap size Δ and the g-factor are substantially underestimated. In particular, owing to the larger nonlinearity at the $n_s/2$ state, its g-factor is underestimated by a factor of about three by using the larger excitation. Therefore, one should keep these nonlinear effects in mind when doing temperature-dependent measurements on such resistive states to obtain accurate results.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

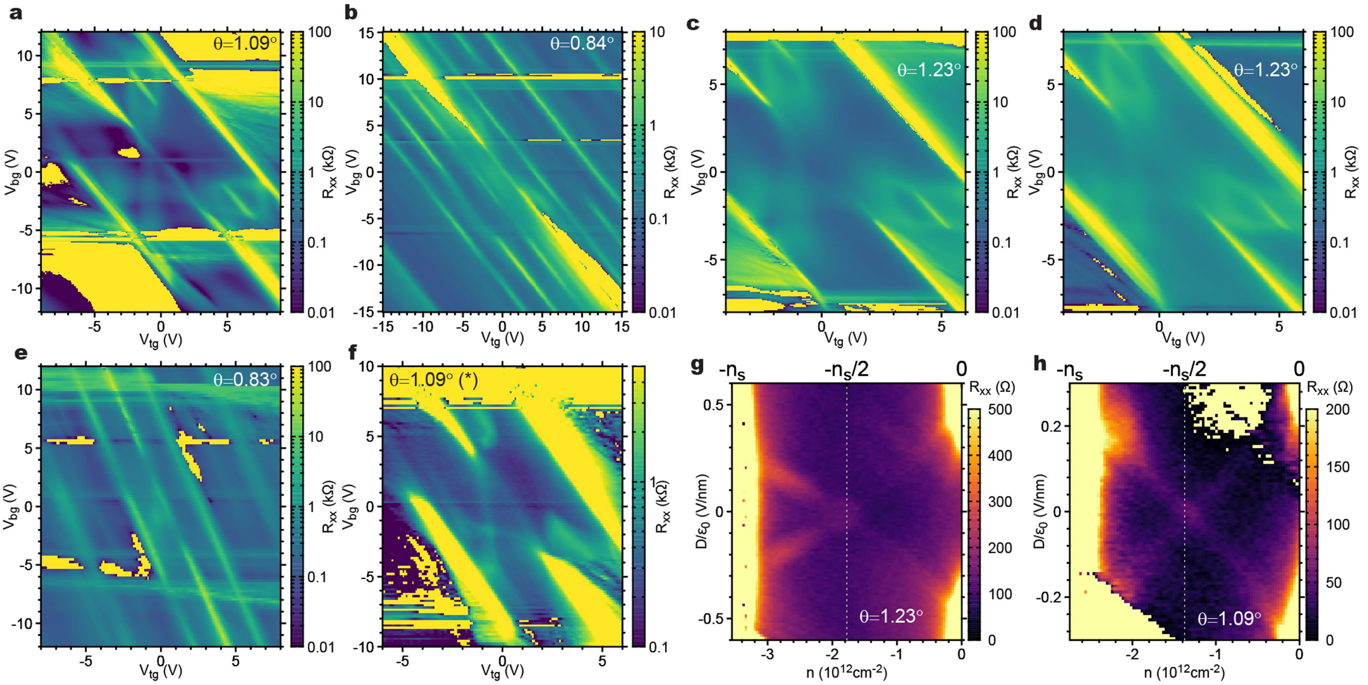
30. Pizzocchero, F. et al. The hot pick-up technique for batch assembly of van der Waals heterostructures. *Nat. Commun.* **7**, 11894 (2016).
31. Purdie, D. G. et al. Cleaning interfaces in layered materials heterostructures. *Nat. Commun.* **9**, 5387 (2018).
32. Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).

Acknowledgements We acknowledge discussions with S. Todadri, L. Fu, P. Kim, X. Liu, S. Fang and E. Kaxiras. This work was supported by the National Science Foundation under award DMR-1809802 (data analysis by Y.C.), the Center for Integrated Quantum Materials under NSF grant DMR-1231319 (fabrication by D.R.-L.), the US DOE, BES Office, Division of Materials Sciences and Engineering under award DE-SC0001819 (*g*-factor analysis by J.M.P.), and the Gordon and Betty Moore Foundation's EPiQS Initiative through grant GBMF4541 to P.J.-H. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by the MEXT, Japan, A3 Foresight by JSPS and the CREST (JPMJCR15F3), JST. This work made use of the Materials Research Science and Engineering Center Shared Experimental Facilities supported by the National Science Foundation (DMR-0819762) and of Harvard's Center for Nanoscale Systems, supported by the NSF (ECS-0335765). D.R.-L. acknowledges partial support from Fundació Bancaria "la Caixa" (LCF/BQ/AN15/10380011) and from the US Army Research Office grant number W911NF-17-S-0001 (measurements). O.R.-B. acknowledges support from Fundació Privada Cellex.

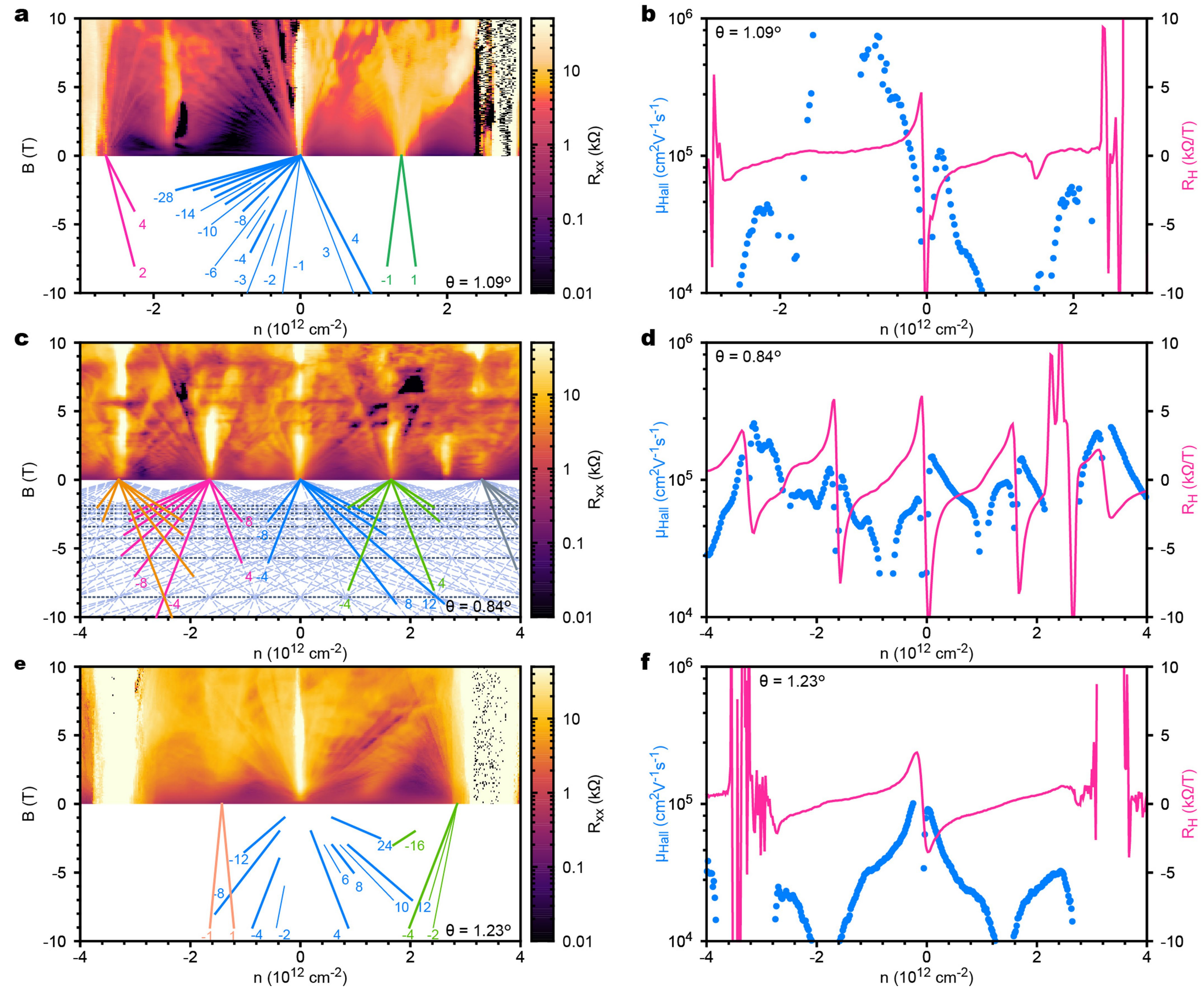
Author contributions Y.C., D.R.-L., O.R.-B. and J.M.P contributed to sample fabrication and transport measurements. Y.C., D.R.-L., O.R.-B., J.M.P and P.J.-H. performed data analysis. K.W. and T.T. provided h-BN samples. Y.C., D.R.-L., O.R.-B., J.M.P. and P.J.-H. wrote the manuscript with input from all co-authors.

Competing interests The authors declare no competing interests.

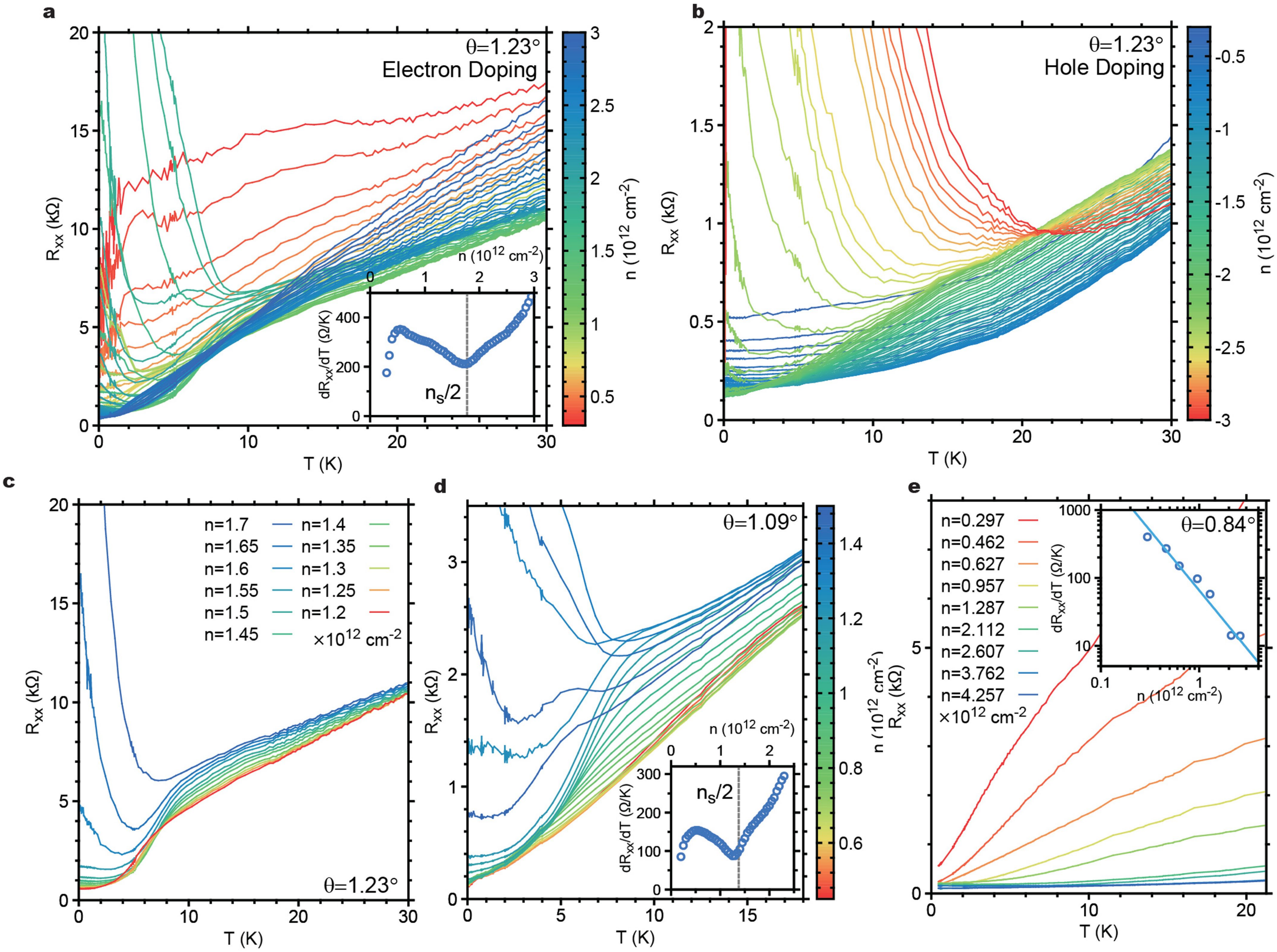
Additional information
Correspondence and requests for materials should be addressed to Y.C. or P.J.-H.
Peer review information *Nature* thanks Ming-Hao Liu, Hu-Jong Lee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



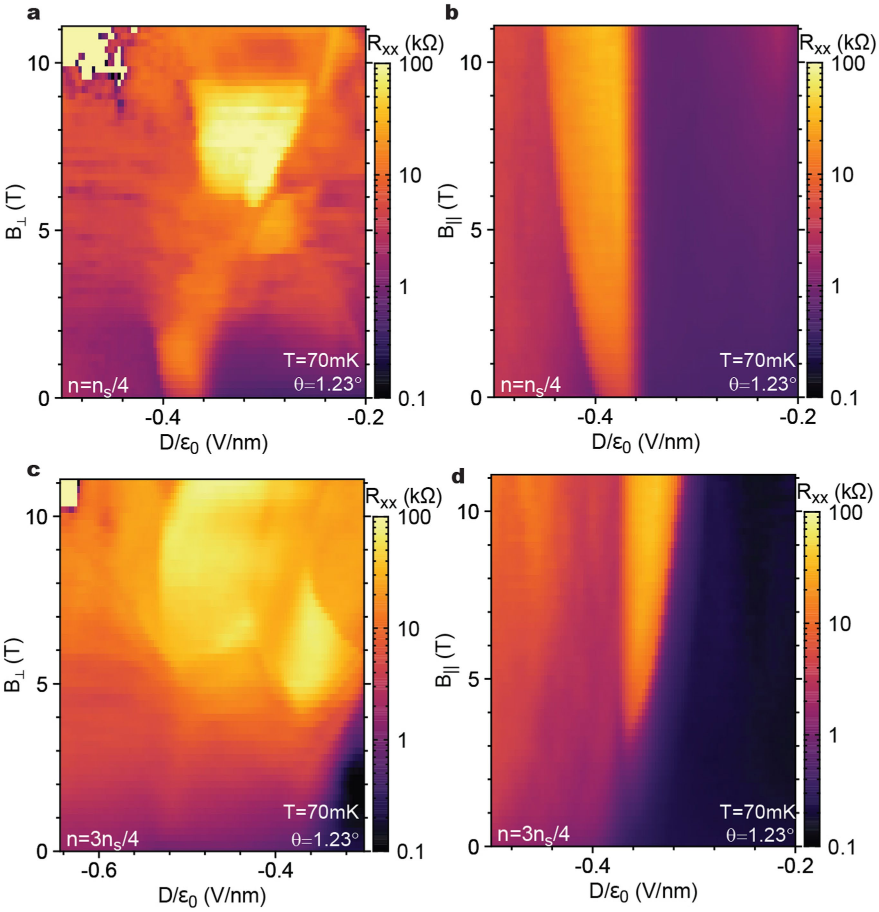
Extended Data Fig. 1 | V_{tg}–V_{bg} resistance maps of measured TBBG devices. **a**–**f**, Resistance versus V_{tg} and V_{bg} for the six TBBG devices measured, which correspond to the six rows shown in Extended Data Table 1, respectively. **g**, **h**, Cross-like feature near $-n_s/2$ in TBBG samples with twist angles $\theta = 1.23°$ (**g**) and $\theta = 1.09°$ (**h**), which might signal the onset of a correlated state.



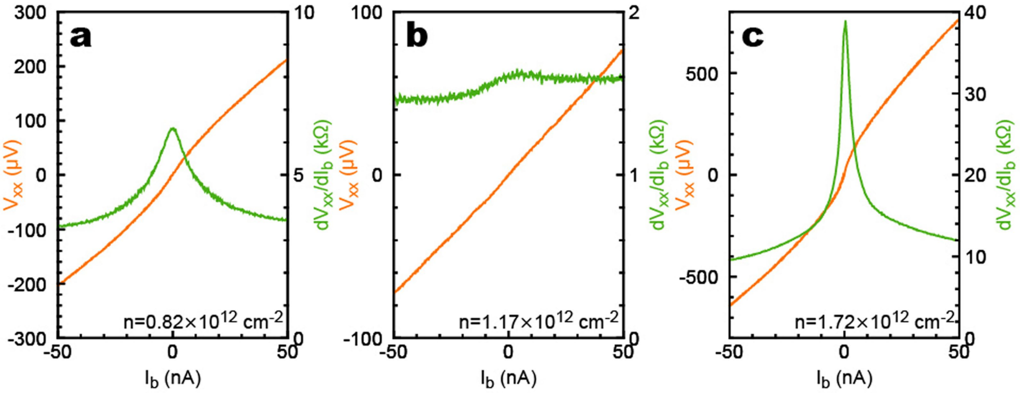
Extended Data Fig. 2 | Landau fan diagrams and Hall mobilities of the TBBG devices. a, Resistance of the 1.09° sample versus carrier density and perpendicular magnetic field. **b**, Hall mobility μ_{Hall} (left axis) and Hall coefficient R_H (right axis) in the 1.09° sample at different carrier densities. **c**–**f**, Same measurements as in **a**, **b** but for the 0.84° (**c**, **d**) and 1.23° (**e**, **f**) samples, respectively. All measurements are taken at $T < 100$ mK. The data for the 1.09° device are taken at $D/\varepsilon_0 = 0.2$ V nm^{-1} while the data for the other two devices are taken at $D = 0$.



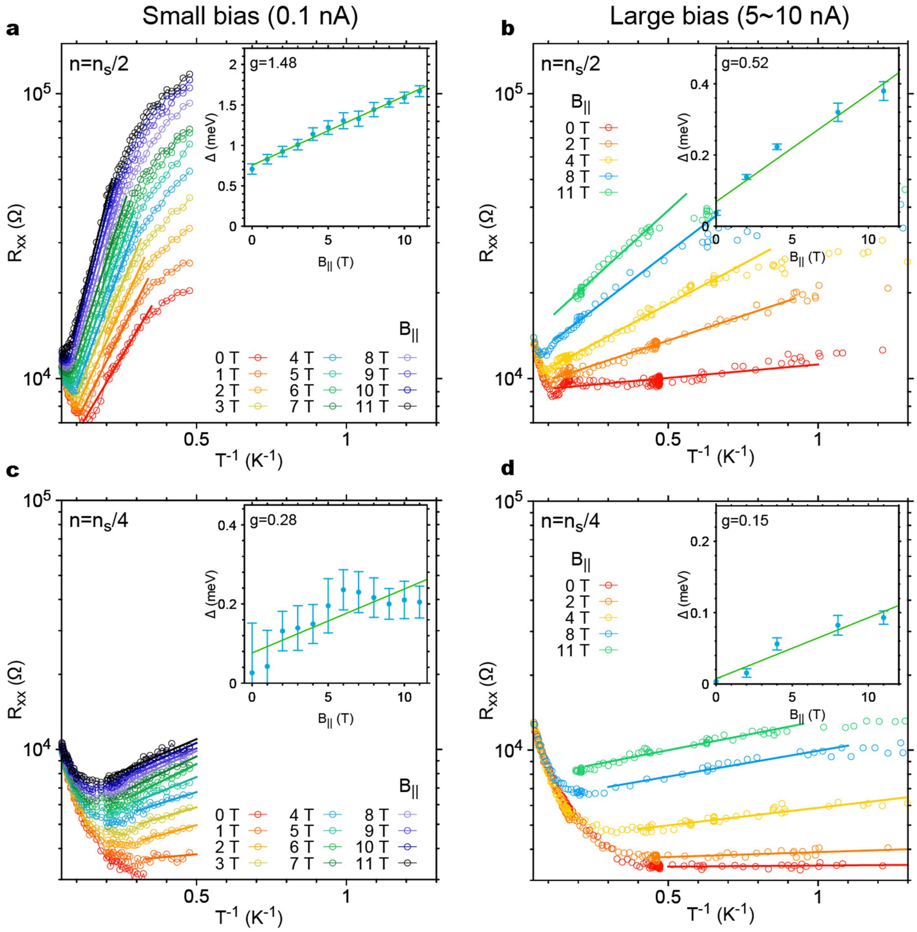
Extended Data Fig. 3 | Linear resistance versus temperature behaviour in TBBG. **a**, **b**, Resistance versus temperature curves at different charge densities in the 1.23° sample for the electron-doping side (**a**) and the hole-doping side (**b**). The inset in **a** shows the slope dR_{xx}/dT of the linear R–T behaviour as a function of n for $T > 10$ K. **c**, Selected R–T curves near $n_s/2$ from **a**. **d**, Similar linear R–T behaviour in the 1.09° device. The inset shows the slope dR_{xx}/dT. **e**, Density-dependent sublinear/linear R–T behaviour in the 0.84° device. The inset shows the slope dR_{xx}/dT versus n in log–log scale. The slope is proportional to n to the power of −1.77.



Extended Data Fig. 4 | Additional magnetic field response of TBBG devices. a–d, Response of the $n_s/4$ (**a**, **b**) and $3n_s/4$ (**c**, **d**) states in perpendicular magnetic field (**a**, **c**) and in-plane magnetic field (**b**, **d**) for the θ =1.23° device.



Extended Data Fig. 5 | *I*–*V* curves in the 1.23° TBBG device at different carrier densities. $D/\varepsilon_0 = -0.38$ V nm^{-1}. **a**–**c**, The densities correspond approximately to the $n_s/4$ (**a**) and $n_s/2$ (**c**) insulating states while the density for **b** lies between them. The left axis is the longitudinal voltage V_{xx} and the right axis is the differential resistance dV_{xx}/dI_b.



Extended Data Fig. 6 | Comparison of the gap sizes and the g-factor using small and large excitations. a, **b**, The Arrhenius fits of the resistance at the $n_s/2$ state of the 1.23° TBBG device in an in-plane magnetic field. **c**, **d**, The same fits for the $n_s/4$ state. **a** and **c** are measured using a current excitation of 0.1 nA, while **b** and **d** are measured using a voltage excitation of around 100 μV, which induces a current of around 5–10 nA in the sample. The insets in each panel show the corresponding *g*-factor fittings. In general, by using an excessive excitation, both the energy gaps and the *g*-factor will be underestimated.

Extended Data Table 1 | List of TBBG devices discussed in the main text and Extended Data figures

θ(°)	c_{tg}(F/m^2)	c_{bg}(F/m^2)	$V_{tg,0}$(V)	$V_{bg,0}$(V)	n_s(cm^{-2})
1.09	6.63×10^{-4}	5.02×10^{-4}	0.30	0.58	2.75×10^{12}
1.23	1.06×10^{-3}	7.14×10^{-4}	0.41	-0.04	3.55×10^{12}
1.23	1.06×10^{-3}	7.14×10^{-4}	0.41	-0.04	3.55×10^{12}
0.84	6.87×10^{-4}	6.38×10^{-4}	0.06	0.08	1.65×10^{12}
0.79	1.06×10^{-3}	3.57×10^{-4}	0.18	0.67	1.45×10^{12}
1.09(*)	1.03×10^{-3}	5.12×10^{-4}	0.28	0.45	2.75×10^{12}

The last device is marked with an asterisk to differentiate it from the first device, which happens to have the same twist angle, but it is a totally independent device fabricated on a separate chip.

Tunable spin-polarized correlated states in twisted double bilayer graphene

https://doi.org/10.1038/s41586-020-2458-7

Received: 25 March 2019

Accepted: 19 May 2020

Xiaomeng Liu[1,3]✉, Zeyu Hao[1], Eslam Khalaf[1], Jong Yeon Lee[1], Yuval Ronen[1], Hyobin Yoo[1], Danial Haei Najafabadi[1], Kenji Watanabe[2], Takashi Taniguchi[2], Ashvin Vishwanath[1] & Philip Kim[1]✉

Reducing the energy bandwidth of electrons in a lattice below the long-range Coulomb interaction energy promotes correlation effects. Moiré superlattices—which are created by stacking van der Waals heterostructures with a controlled twist angle[1–3]—enable the engineering of electron band structure. Exotic quantum phases can emerge in an engineered moiré flat band. The recent discovery of correlated insulator states, superconductivity and the quantum anomalous Hall effect in the flat band of magic-angle twisted bilayer graphene[4–8] has sparked the exploration of correlated electron states in other moiré systems[9–11]. The electronic properties of van der Waals moiré superlattices can further be tuned by adjusting the interlayer coupling[6] or the band structure of constituent layers[9]. Here, using van der Waals heterostructures of twisted double bilayer graphene (TDBG), we demonstrate a flat electron band that is tunable by perpendicular electric fields in a range of twist angles. Similarly to magic-angle twisted bilayer graphene, TDBG shows energy gaps at the half- and quarter-filled flat bands, indicating the emergence of correlated insulator states. We find that the gaps of these insulator states increase with in-plane magnetic field, suggesting a ferromagnetic order. On doping the half-filled insulator, a sudden drop in resistivity is observed with decreasing temperature. This critical behaviour is confined to a small area in the density–electric-field plane, and is attributed to a phase transition from a normal metal to a spin-polarized correlated state. The discovery of spin-polarized correlated states in electric-field-tunable TDBG provides a new route to engineering interaction-driven quantum phases.

Moiré superlattices of two-dimensional (2D) van der Waals (vdW) materials provide a new scheme for creating correlated electronic states. By controlling the twist angle θ between atomically thin vdW layers, the size of the moiré unit cell can be tuned[1–3]. In particular, in twisted bilayer graphene (TBG), the weak interlayer coupling can open up energy gaps at the boundary of the mini-Brillouin zone, which modifies the energy bands of the coupled system. Theoretically, it has been predicted that around $\theta \approx 1.1°$ (the so-called magic angle, MA), the interlayer hybridization induces isolated flat bands with drastically reduced bandwidth and enhanced density of states[12]. The combination of flat bands and moiré periodic potential fosters an environment where strongly correlated states can emerge. Recent experiments performed in MA-TBG indeed confirmed the appearance of correlated insulator states associated with the flat bands[4]. Intriguingly, on doping the half-filled insulator, superconductivity was discovered. The phase diagram of MA-TBG thus phenomenologically resembles that of high-temperature superconductors, whose undoped parent compounds are Mott insulators[13]. As a result, there is hope that MA-TBG could be a gateway to understanding the long-lasting puzzle of high-temperature superconductivity. Yet, in recent studies, the connection between superconductivity and the correlated insulator state has been debated[14–19].

One method to study the MA-TBG system is to tune the band structure through the flat-band condition and observe how the correlated physics changes. So far, such experimental control has largely been achieved by fabricating samples with different twist angles. However, different samples—owing to differences in uncontrollable factors such as the alignment with hexagonal boron nitride (hBN), strain and dielectric thickness—often yield contradicting results regarding where the correlated insulator and superconductivity appear. Only limited tunability has been demonstrated in TBG by the application of hydrostatic pressure[6]. In ABC trilayer graphene/hBN superlattices, the electric field has been shown to modulate the correlated insulator gap[9], opening up the possibility of continuous tuning of the moiré flat band with electric field. However, the difficulty in identifying and preserving the unstable ABC trilayer graphene, together with the precise alignment required between the graphene and hBN layers, makes it a less accessible platform. Here we demonstrate a wide range of electric field tunability in the moiré flat band of twisted double bilayer graphene (TDBG),

---

[1]Department of Physics, Harvard University, Cambridge, MA, USA. [2]National Institute for Material Science, Tsukuba, Japan. [3]These authors contributed equally: Xiaomeng Liu, Zeyu Hao. ✉e-mail: xiaomeng@princeton.edu; pkim@physics.harvard.edu

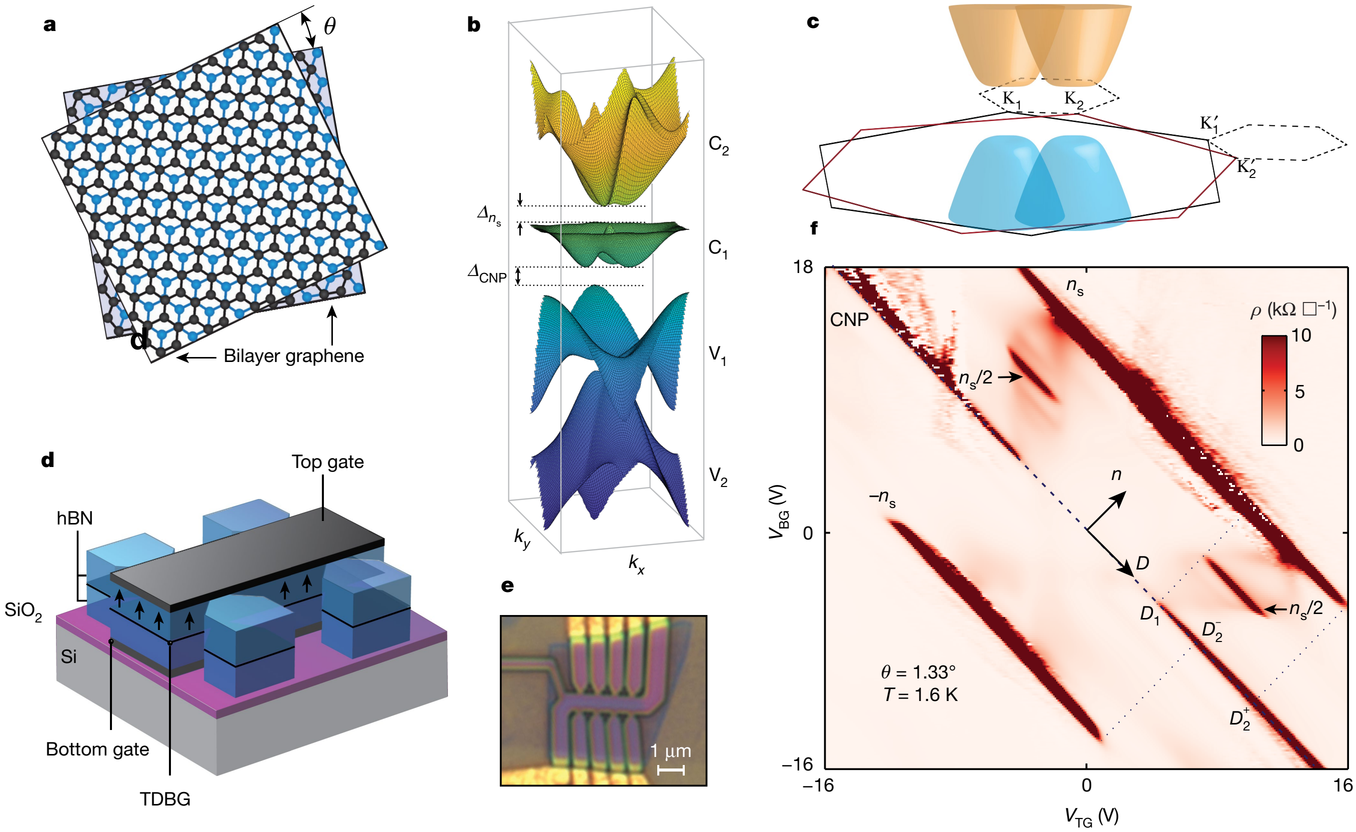


Fig. 1 | Band structure and insulating states in the $\theta = 1.33°$ sample. **a**, Schematic of TDBG with a twist angle θ. **b**, Calculated band structure for $\theta = 1.33°$ TDBG at an optimal displacement field. k_x and k_y are wave vectors in the x and y direction. **c**, Brillouin zone and band structure of the two individual bilayer graphene layers under a perpendicular displacement field. The dashed hexagons represent the mini-Brillouin zone of the moiré superlattice. K_1 and K_1' (K_2 and K_2') are the two valleys of the top (bottom) bilayer graphene. **d**, Device schematic with graphite top and bottom gates. The arrows represent displacement fields generated by the top gate. **e**, Optical microscope image of the 1.33° device. **f**, Resistivity as a function of top and bottom gate voltages. CNP, full-filled gaps ($\pm n_s$) and half-filled gaps ($n_s/2$) are marked. The displacement field where the CNP starts to open up a gap is labelled with D_1 and $D_2^{\pm}$ labels where the gap at the full electron (hole) band filling closes.

consisting of two Bernal-stacked bilayer graphene sheets misaligned with a twist angle θ (Fig. 1a).

In twisted systems, the twist angle for achieving a flat band is determined by the band structure of the individual layer and the interlayer coupling strength. Unlike monolayer graphene, the band structure of Bernal-stacked bilayer graphene can be tuned by a perpendicular displacement field D (ref. [20]). As $|D|$ increases, the parabolic band touching at charge neutrality of bilayer graphene opens up a gap and the bottom (top) of the conduction (valence) band lifts up (down) into a shallow Mexican-hat-shaped energy dispersion distorted by trigonal warping[21]. The gap in bilayer graphene can be as large as 200 meV for large $|D|$ before the gate dielectric breaks down[22]. In TDBG, where two bilayers are stacked, the displacement field affects the energy dispersion of each constituent bilayer graphene, allowing a new experimental 'knob' to tune the flat-band condition (Fig. 1c). Figure 1b shows moiré band structures calculated at finite D using the single-particle continuum model approximation[12,23–27]. We find that a well-isolated narrow conduction band can appear for a range of twist angles θ, where the interband energy gaps and bandwidth can be controlled by the displacement field (see Methods for details).

We fabricated TDBG devices by tearing and stacking Bernal-stacked bilayer graphene[28,29]. We measured in total seven devices with twist angle $\theta = 1.26$, 1.32, 1.33, 1.41, 1.48, 1.53 and 2.00°, with the first six devices showing signatures of correlation effects. All of the devices measured are encapsulated by hBN. Top gates are made from graphite or metal, and bottom gates are made from graphite or silicon (details for each device structure are shown in Extended Data Fig. 9). We focus our study on the two representative devices $\theta = 1.33°$ and $\theta = 1.26°$, and summarize the behaviour of the other devices in Methods and Extended Data Table 1. The top and bottom gates with voltages V_{TG} and V_{BG} are used to control the density of electrons, n, and displacement field, D, independently: $n = (C_{\mathrm{TG}}V_{\mathrm{TG}} + C_{\mathrm{BG}}V_{\mathrm{BG}})/e$ and $D = (C_{\mathrm{TG}}V_{\mathrm{TG}} - C_{\mathrm{BG}}V_{\mathrm{BG}})/2$, where C_{TG} (C_{BG}) is the capacitance between the TDBG and the top (bottom) gate and e is the elementary charge.

Figure 1f shows the four-probe resistivity ρ measured in the TDBG with $\theta = 1.33°$ as a function of V_{TG} and V_{BG} at temperature $T = 1.6$ K. CNP represents the charge neutral point of the TDBG and n_s denotes the full filling of the flat band, corresponding to four electrons per moiré unit cell, originating from the spin and valley degeneracy. For a linecut along a constant displacement field $D \approx (D_1 + D_2^-)/2$ (the positions of D_1 and $D_2^{\pm}$ are labelled in Fig. 1f), ρ shows several insulating states where the corresponding conductance $\sigma = \rho^{-1}$ vanishes as the temperature T decreases (Fig. 2a), suggesting a gap opening at the Fermi level of the system. Some insulating regions identified in Fig. 1f can be well explained by the single-particle band structure presented in Fig. 1b. For example, we find that the CNP is gapless at $D = 0$ but develops a gap for $|D| > D_1 \neq 0$. Similarly, at full moiré band filling $n = \pm n_s$, energy gaps $\Delta_{\pm n_s}$ are present within displacement field ranges $|D| < D_2^{\pm}$. Consequently, for $D_1 < |D| < D_2^+$ ($D_1 < |D| < D_2^-$), there is an isolated conduction (valence) band. Note that $D_2^{\pm}$ is different in the conduction band (+) and valence band (−), owing to the lack of electron–hole symmetry in TDBG. All these single-particle bandgaps are nicely captured by our calculation based on a continuum model (Extended Data Fig. 1). The calculation also captures the cross-like feature in Fig. 1f, which matches with the

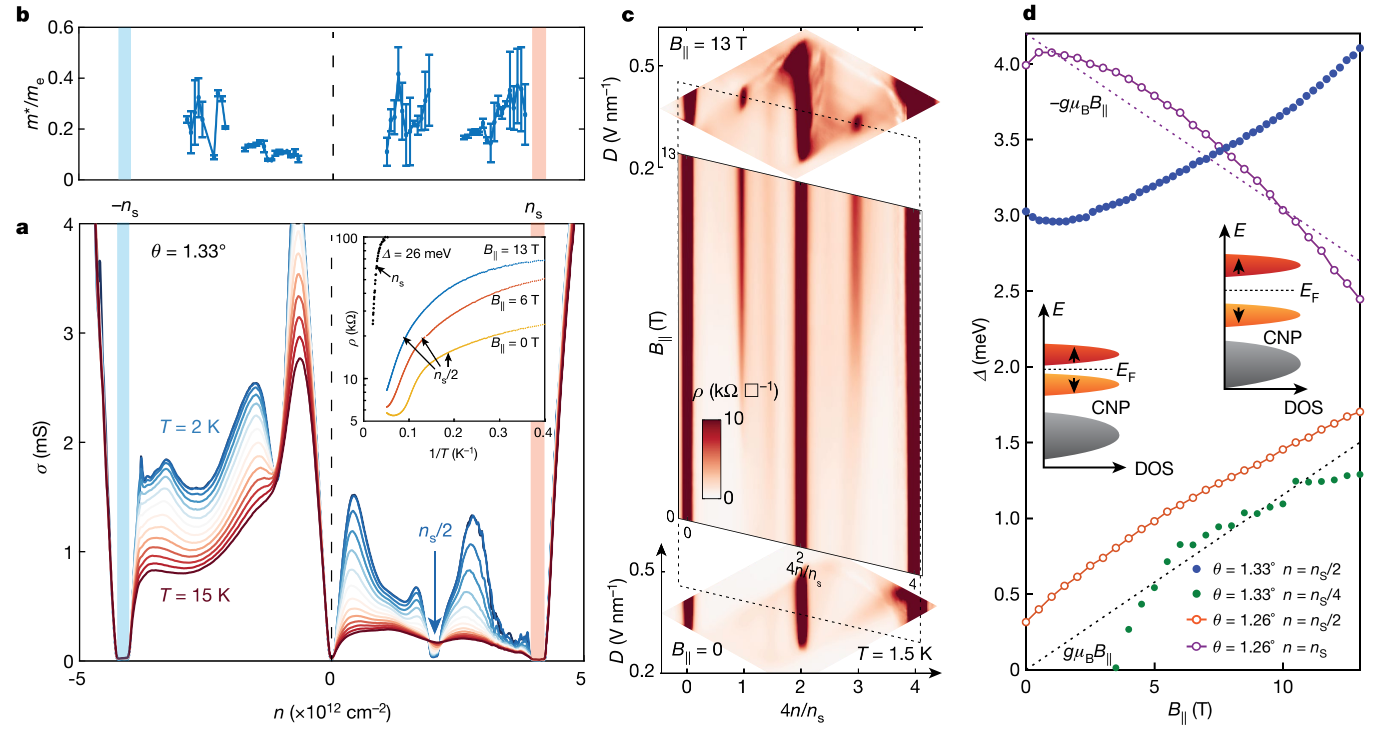


Fig. 2 | Spin polarization of the correlated insulator states. a, Temperature dependence of conductivity (σ) as a function of carrier density at a constant displacement field that passes through the half-filled insulator ($D \approx (D_1 + D_2^-)/2$). Inset: Arrhenius plot for the full-filled insulating state (n_s) and the half-filled insulating state ($n_s/2$) under different in-plane magnetic fields. **b**, Effective mass measured by temperature-dependent quantum oscillations corresponding to **a**. **c**, Development of the correlated insulating states with in-plane magnetic fields. Top and bottom panels compare resistivity as a function of n and D under in-plane magnetic fields of $B_{\parallel} = 0$ T (bottom) and $B_{\parallel} = 13$ T (top). The middle panel shows the continuous evolution of the correlated states by taking a linecut along the dashed lines in the top and bottom panels. All three panels are measured at a temperature of $T = 1.5$ K. **d**, Half-filled insulating gap $\Delta_{n_s/2}$, quarter-filled insulating gap $\Delta_{n_s/4}$ and full-filled gap Δ_{n_s} as a function of in-plane magnetic field. The black dashed line indicates Zeeman energy with $g = 2$. $\Delta_{n_s/2}$ of both devices increases with in-plane magnetic field, indicating spin polarization of the half-filled insulator. We also note that the single-particle gap Δ_{n_s} between c_1 and c_2 (purple curve) decreases linearly with Zeeman energy with a g-factor of 2 (purple dashed line). Insets: schematic of the half-filled insulating state at zero (left) and large (right) in-plane fields. The x axis is density of state (DOS) and the y axis is energy (E_F is the Fermi energy). Grey represents the inert valance band while orange (red) represents the lower (upper) half of the first conduction band.

van Hove singularities of the bands (details in Methods). Lastly, the calculated band structure (Fig. 1b) indeed demonstrates the existence of an isolated flat band at $\theta = 1.33°$ under finite displacement field with a bandwidth around 10–15 meV.

In this single-particle band structure, we expect a narrow but uninterrupted spectrum within the lowest moiré conduction band (c_1), separated by bandgaps from both the valence band (v_1) and higher conduction band (c_2) for $D_1 < |D| < D_2^-$. However, we observe the development of a well-defined insulating behaviour at half-filling $n = n_s/2$ (Figs. 1f, 2a). The onset displacement field of this insulating state coincides with D_1. However, it ends well before D reaches D_2^+, suggesting both the isolation and the flatness of the band are required for creating the observed correlated gap (Fig. 1f). Along the same linecut shown in Fig. 2a ($D \approx (D_1 + D_2^-)/2$), we measure the effective cyclotron mass m^* from the temperature-dependent magnetoresistance oscillations (Extended Data Fig. 8). Figure 2b shows that $m^* \approx 0.2m_e$ for the first valence band (v_1) and $m^* \approx 0.3m_e$ for the first conduction band (c_1), where m_e is the bare electron mass. Considering the effective mass of Bernal-stacked bilayer graphene is about $0.04m_e$ (ref. [30]), the experimentally observed large m^* indicates an order of magnitude narrower bandwidth than that of bilayer graphene bands folded in the moiré superlattice Brillouin zone, especially for the c_1 band. We then use the conduction band effective mass $m^* = 0.3m_e$ to estimate the bandwidth of the c_1 band to be about 10 meV. This bandwidth matches with the continuum model calculation of TDBG[23], confirming the existence of the flat band experimentally. The absence of correlated insulating behaviour in the hole-doped regime under similar experimental conditions can be explained by the larger bandwidth of the moiré valence band v_1 than that of c_1 (Methods).

We measure the size of the insulating gaps from the activating behaviour of ρ (Fig. 2a, inset). For $\theta = 1.33°$ TDBG, the half-filled insulator is robust with an energy gap of $\Delta_{n_s/2} = 3$ meV and persists up to a perpendicular magnetic field $B_{\perp} \approx 7$ T (Extended Data Fig. 7). As the c_1 band is spin and valley degenerate in a single-particle picture, the half-filled insulator is probably polarized in the fourfold spin–valley space. The in-plane magnetic field $B_{\parallel}$ can be used to probe the spin structure of the state without substantially coupling to the valley degrees of freedom in the regime where in-plane orbital effect is negligible. In MA-TBG, it has been shown that $B_{\parallel}$ reduces $\Delta_{n_s/2}$. Figure 2a inset and Fig. 2c show the change of ρ as a function of $B_{\parallel}$ in our TDBG sample. We find that the half-filled insulator becomes more insulating as $B_{\parallel}$ increases (Fig. 2a, inset) and the displacement field range spanned by the half-filled insulator expands (Fig. 2c). More quantitatively, we find that the growth of $\Delta_{n_s/2}$ roughly follows the Zeeman energy scale $g\mu_B B_{\parallel}$, where μ_B is the Bohr magneton and the effective g-factor $g = 2$ (dashed black line in Fig. 2d). This observation is consistent with a picture where the occupied states (half of the states in c_1) are spin polarized along the direction of the external magnetic field. The unoccupied excited states then carry the opposite spin, separated by a ferromagnetic gap due to spontaneous symmetry breaking at half-filling. For spin-1/2, the Zeeman term lowers the energy of the filled states $\Delta E_{\downarrow} = -g\mu_B B/2$, while boosting the energy of the empty states with opposite spins by $\Delta E_{\uparrow} = g\mu_B B/2$, pushing the two bands further apart and enhancing the gap (as illustrated by Fig. 2d, insets). Calculations from the Hartree–Fock

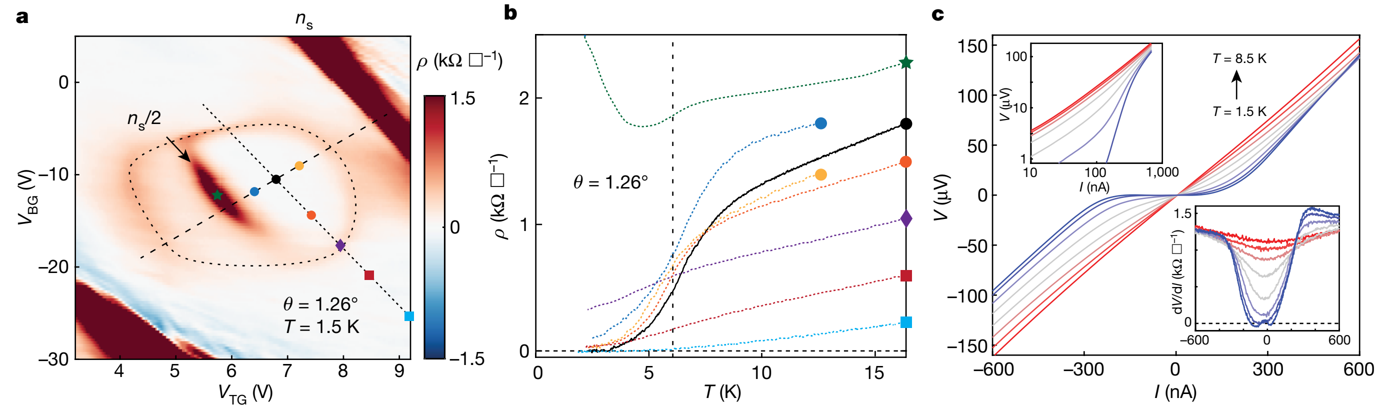


Fig. 3 | Critical transition behaviour in the $\theta = 1.26°$ sample. a, Resistivity map around the half-filled insulator. Dashed circle marks the halo. **b**, Resistivity as a function of temperature at different spots marked by the coloured symbols in **a** using the corresponding colours and shapes. **c**, I–V curves at the black circle in **a**. Top left inset: I–V on a logarithmic scale, demonstrating the BKT-like power-law behaviour (see Extended Data Fig. 4 for more details). Bottom right inset: dV/dI as a function of bias current, which shows a critical current of about 300 nA. The curves cover temperature from 1.5 K (blue) to 8.5 K (red).

approximation also support the existence of a spin-polarized correlated insulating state at half-filling in TDBG[23,24].

In the $\theta = 1.33°$ device, applying $B_{\|}$ also induces additional correlated insulating states at quarter-filling ($n = \frac{1}{4}n_s$) and three-quarter-filling ($n = \frac{3}{4}n_s$) (Fig. 2c). The quarter-filled insulating gap opens at $B_{\|} \approx 4$ T and increases as $B_{\|}$ increases (Fig. 2d). According to the hierarchy of the symmetry-broken states within mean-field theory[23], the quarter-filled gaps separate the ground state and the excited state of the same spin and opposite valleys, and thus should be relatively insensitive to in-plane magnetic fields. However, the enhancement of quarter-filled gaps with $B_{\|}$ and the positions where quarter-filled insulating states appear in the n–D plane (Extended Data Fig. 2) suggest that these gaps probably separate states of opposite spin, hinting that the origin of these strongly correlated states goes beyond a simple mean-field approach.

In the $\theta = 1.26°$ sample, a similar spin-polarized half-filled insulating state is observed (Fig. 3a), with a much smaller correlated gap $\Delta_{n_s/2} = 0.3$ meV (red line in Fig. 2d). On doping the half-filled insulator, we identify the appearance of a 'halo' (marked with a dashed circle in Fig. 3a) surrounding the half-filled insulating state in the V_{TG}–V_{BG} plane. On the halo, the resistivity is slightly higher than that of the nearby region. Such a halo-like feature commonly appears around the correlated insulating states in different samples with varying twist angles (Fig. 4a–d, Extended Data Fig. 9). The half-filled insulting state divides the halo-like region into two. For the samples with a strong half-filled insulating gap, Hall measurements performed at low magnetic fields (Extended Data Fig. 2) show a sign change of the Hall signal across the boundary of the halo and also across the correlated insulator. A similar observation has been noted in a recent related study[31]. The sign of the Hall signal inside the halo complies with the carrier concentration counted from half-filling. This suggests the metallic state in the halo is obtained by adding carriers to the spin-polarized band at half-filling while retaining the spin-splitting of the band (Fig. 2d, inset), and therefore is probably a ferromagnetic metal. As the Hall signal outside the halo matches the expectation for a moiré band without correlation, the halo marks the border between the spin-polarized and the spin-unpolarized metallic states (Extended Data Fig. 2c).

Studying the temperature dependence of the resistivity, $\rho(T)$, inside the halo, we identify a critical transition with a sudden drop in resistivity as the temperature decreases. Figure 3b shows the resistivity measured at different gate configurations marked by matching symbols in Fig. 3a. We note that the critical transition behaviour, namely the sudden drop in resistivity, occurs only inside the halo. In contrast, resistivity outside the halo increases linearly with temperature. The resistivity behaviour outside the halo is most likely due to ballistic transport at low temperatures and enhanced phonon scattering at elevated temperatures. The critical transition behaviour of $\rho(T)$ inside the halo, however, appears non-trivial. The $\rho(T)$ curve of the 1.26° device (black curve in Fig. 3b) in particular strongly resembles that of a superconductor, with near-zero resistivity below 3.5 K. The current–voltage (I–V) curve also shows superconducting-like nonlinear behaviours: dV/dI vanishes for bias current smaller than the critical current, $I < I_c$, and increases to a near-constant value that is close to the normal resistivity above the critical transition for $I > I_c$ (Fig. 3c, bottom right inset). This nonlinear I–V characteristic is distinct from that of a heating effect (see Methods for a more detailed analysis) and seemingly follows that of the Berezinskii–Kosterlitz–Thouless (BKT) transition (Extended Data Fig. 4e).

While $\rho(T)$ and the I–V characteristic discussed above for the 1.26° device are suggestive of superconductivity, we note that several factors require careful consideration. First, we have not observed direct evidence of superconducting phase coherence, such as the Fraunhofer pattern under magnetic fields. Second, $\rho(T \ll T_c) \approx 0$ has been observed only for the 1.26° device. Figure 4a–d shows four other devices we measured with the twist angle ranging between 1.32° and 1.48°. In these devices, similar to the 1.26° device, critical transition behaviours in the $\rho(T)$ curves are commonly observed inside the halo region that surrounds the half-filled insulator. These critical behaviours are best illustrated by the clear peaks in dρ/dT, which are absent outside the halo (Fig. 4e, f). The critical temperatures, defined as the temperature where dρ/dT is maximum, are similar across all devices ($T_c = 6$–9 K), despite their very different half-filled insulating gap sizes (Extended Data Table 1). However, the low-temperature resistivity $\rho(T \ll T_c)$ does not reach zero, unlike in the 1.26° device (Fig. 4e, Extended Data Table 1). Strong nonlinear I–V at low temperatures is also absent in these devices.

On the basis of these experimental findings, we propose a few scenarios to explain the observed critical transition behaviour. One possibility is that the critical transition is a result of Cooper-pair formation, but superconductivity is developed only in the 1.26° device. In other devices, the establishment of phase coherence may be inhibited by an inhomogeneous distribution of strains or disorder. Alternatively, the critical transition may correspond to a ferromagnetic transition of the doped half-filled insulating states. Here we note that the critical transition behaviours occur only inside the halo region, which we associate with ferromagnetic metallic states. As the temperature increases, the ferromagnetic metal turns into a normal metal when the correlation effect vanishes. Below the critical temperature, carrier scattering processes related to spin-flip can potentially be suppressed by the ferromagnetic order, resulting in a reduced resistivity. These two scenarios

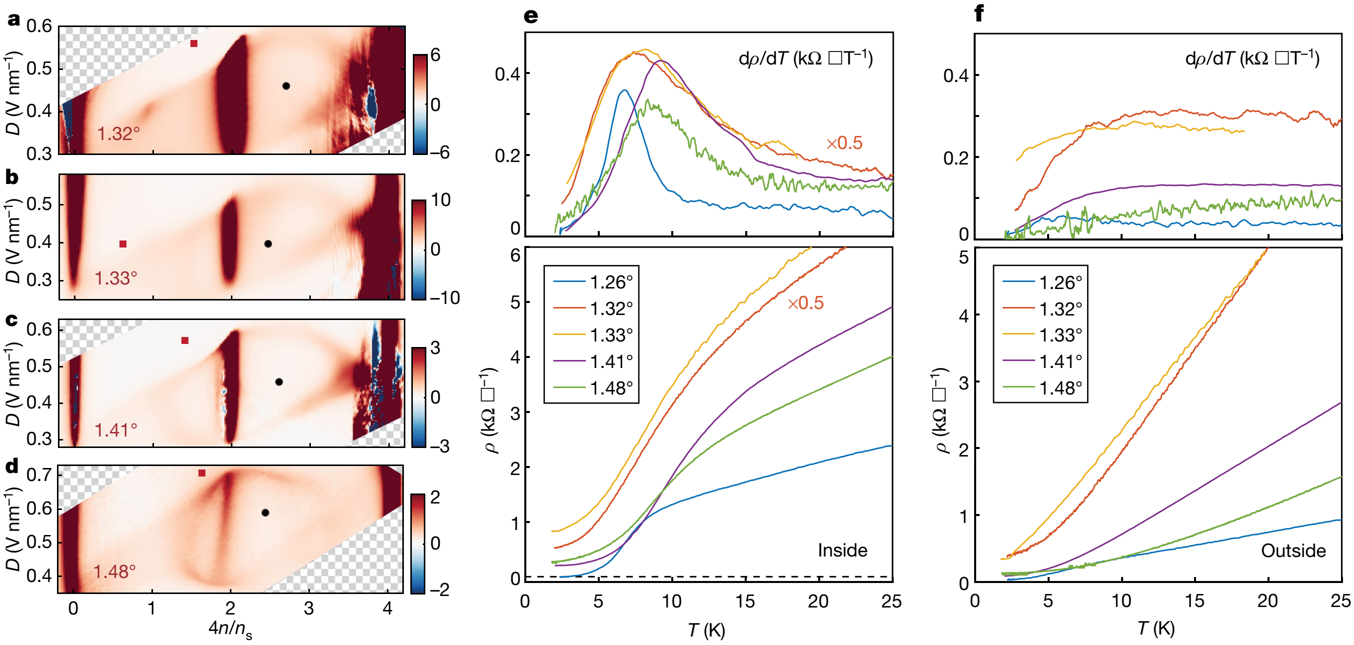


Fig. 4 | Gate-tunable flat band and critical behaviour in a series of twist angles. a–**d**, Resistivity as a function of filling fraction and displacement field around the half-filled insulators measured in four different samples with twist angles of 1.32° (**a**), 1.33° (**b**), 1.41° (**c**) and 1.48° (**d**). **e**, **f**, Resistivity as a function of temperature inside (**e**) and outside (**f**) the halo regions surrounding the correlated insulator. The gate configuration for each curve is marked by the red (outside the halo) and the black (inside the halo) symbols in **a**–**d**. Top panels show the derivative of the resistivity to highlight the critical transition behaviour.

do not necessarily compete with each other, leaving open the possibility of a ferromagnetic superconductor (Methods, Extended Data Fig. 5). The highly tunable electronic structures of TDBG demonstrated here and in related studies[31–34] may provide a new route to engineer correlated phenomena in a moiré superlattice.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2458-7.

1. Dean, C. R. et al. Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
2. Ponomarenko, L. A. et al. Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
3. Hunt, B. et al. Massive Dirac fermions and Hofstadter butterfly in a van der Waals heterostructure. *Science* **340**, 1427–1430 (2013).
4. Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
5. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
6. Yankowitz, M. et al. Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).
7. Sharpe, A. L. et al. Emergent ferromagnetism near three-quarters filling in twisted bilayer graphene. *Science* **365**, 605–608 (2019).
8. Serlin, M. et al. Intrinsic quantized anomalous Hall effect in a moiré heterostructure. *Science* **367**, 900–903 (2020).
9. Chen, G. et al. Evidence of a gate-tunable Mott insulator in a trilayer graphene moiré superlattice. *Nat. Phys.* **15**, 237–241 (2019).
10. Wang, L. et al. Magic continuum in twisted bilayer WSe_2. Preprint at http://arxiv.org/abs/1910.12147 (2019).
11. Regan, E. C. et al. Mott and generalized Wigner crystal states in WSe_2/WS_2 moiré superlattices. *Nature* **579**, 359–363 (2020).
12. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
13. Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. From quantum matter to high-temperature superconductivity in copper oxides. *Nature* **518**, 179–186 (2015).
14. Koshino, M. et al. Maximally localized Wannier orbitals and the extended Hubbard model for twisted bilayer graphene. *Phys. Rev. X* **8**, 031087 (2018).
15. Kang, J. & Vafek, O. Symmetry, maximally localized Wannier states, and a low-energy model for twisted bilayer graphene narrow bands. *Phys. Rev. X* **8**, 031088 (2018).
16. Po, H. C., Zou, L., Vishwanath, A. & Senthil, T. Origin of Mott insulating behavior and superconductivity in twisted bilayer graphene. *Phys. Rev. X* **8**, 031089 (2018).
17. Stepanov, P. et al. The interplay of insulating and superconducting orders in magic-angle graphene bilayers. Preprint at http://arxiv.org/abs/1911.09198 (2019).
18. Saito, Y., Ge, J., Watanabe, K., Taniguchi, T. & Young, A. F. Decoupling superconductivity and correlated insulators in twisted bilayer graphene. Preprint at http://arxiv.org/abs/1911.13302 (2019).
19. Arora, H. S. et al. Superconductivity without insulating states in twisted bilayer graphene stabilized by monolayer WSe_2. Preprint at http://arxiv.org/abs/2002.03003 (2020).
20. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009).
21. McCann, E. & Koshino, M. The electronic properties of bilayer graphene. *Rep. Prog. Phys.* **76**, 056503 (2013).
22. Zhang, Y. et al. Direct observation of a widely tunable bandgap in bilayer graphene. *Nature* **459**, 820–823 (2009).
23. Lee, J. Y. et al. Theory of correlated insulating behaviour and spin-triplet superconductivity in twisted double bilayer graphene. *Nat. Commun.* **10**, 5333 (2019).
24. Zhang, Y.-H., Mao, D., Cao, Y., Jarillo-Herrero, P. & Senthil, T. Nearly flat Chern bands in moiré superlattices. *Phys. Rev. B* **99**, 075127 (2019).
25. Koshino, M. Band structure and topological properties of twisted double bilayer graphene. *Phys. Rev. B* **99**, 235406 (2019).
26. Chebrolu, N. R., Chittari, B. L. & Jung, J. Flat bands in twisted double bilayer graphene. *Phys. Rev. B* **99**, 235417 (2019).
27. Choi, Y. W. & Choi, H. J. Intrinsic band gap and electrically tunable flat bands in twisted double bilayer graphene. *Phys. Rev. B* **100**, 201402 (2019).
28. Cao, Y. et al. Superlattice-induced insulating states and valley-protected orbits in twisted bilayer graphene. *Phys. Rev. Lett.* **117**, 116804 (2016).
29. Kim, K. et al. van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
30. Li, J. et al. Effective mass in bilayer graphene at low carrier densities: the role of potential disorder and electron–electron interaction. *Phys. Rev. B* **94**, 161406 (2016).
31. He, M. et al. Tunable correlation-driven symmetry breaking in twisted double bilayer graphene. Preprint at http://arxiv.org/abs/2002.08904 (2020).
32. Shen, C. et al. Correlated states in twisted double bilayer graphene. *Nat. Phys.* **16**, 520–525 (2020).
33. Cao, Y. et al. Tunable correlated states and spin-polarized phases in twisted bilayer–bilayer graphene. *Nature* https://doi.org/10.1038/s41586-020-2260-6 (2020).
34. Burg, G. W. et al. Correlated insulating states in twisted double bilayer graphene. *Phys. Rev. Lett.* **123**, 197702 (2019).

Methods

Band structure of TDBG

The band structure of TDBG with Bernal-stacked bilayers was obtained as follows. In TDBG, each bilayer graphene has a tight-binding Bloch Hamiltonian at a momentum **k** given by

$$H(\mathbf{k}) = \begin{pmatrix} U_1 + \Delta & -\gamma_0 f(\mathbf{k}) & \gamma_4 f^*(\mathbf{k}) & \gamma_1 \\ -\gamma_0 f^*(\mathbf{k}) & U_1 & \gamma_3 f(\mathbf{k}) & \gamma_4 f^*(\mathbf{k}) \\ \gamma_4 f(\mathbf{k}) & \gamma_3 f^*(\mathbf{k}) & U_2 & -\gamma_0 f(\mathbf{k}) \\ \gamma_1 & \gamma_4 f(\mathbf{k}) & -\gamma_0 f^*(\mathbf{k}) & U_2 + \Delta \end{pmatrix} \qquad (1)$$

which is labelled in the order of A1, B1, A2 and B2 sites of the top (1) and the bottom (2) Bernal stacked bilayer graphene. In function $f(\mathbf{k}) \equiv \sum_l e^{i\mathbf{k}\cdot\delta_l}$, i is the imaginary unit, index l runs from one to three, $\delta_1 = a(0, 1)$, $\delta_2 = a(\sqrt{3}/2, -1/2)$ and $\delta_3 = a(-\sqrt{3}/2, -1/2)$, with $a = 1.42$ Å. $f^*(\mathbf{k})$ is the complex conjugate of $f(\mathbf{k})$. In particular, the electrostatic energy difference U between the top and bottom layers is an important tuning parameter controlled by D. With this Hamiltonian, one can follow the continuum model approach in ref. [12] to calculate the moiré band structure.

In the numerical simulation, we use phenomenological parameters

$$(\gamma_0, \gamma_1, \gamma_3, \gamma_4, \Delta) = (2{,}610,\ 361,\ 283,\ 138,\ -15)\ \text{meV} \qquad (2)$$

obtained from ref. [35]. Compared with TBG, TDBG has additional parameters γ_1 and γ_3 (trigonal warping), γ_4 (particle-hole asymmetry) and Δ, in addition to γ_0 (nearest-neighbour hopping). Here, γ_1 and Δ are the interlayer hopping and the on-site energy at A–B stacked sites, where the A site of the first layer (A1) sits on top of the B site of the second layer (B2), respectively. Although these parameters are much smaller than γ_0, they are important to understand the experimental data. In particular, for vanishing U, a finite value of γ_3 yields a larger bandwidth and overlap between c_1 and v_1. This is why the system is metallic at the CNP and there is no magic-angle condition at $D = 0$. Furthermore, γ_4 and Δ give rise to the electron–hole asymmetry. Owing to these terms, the bandwidth of v_1 is much larger than that of c_1, resulting in smaller band isolation for v_1 (Extended Data Fig. 1). For the moiré hopping parameters, $(w_0, w_1) = (0.08, 0.1)$ eV is used to account for the relaxation effect described by ref. [36]. The relaxation increases the gap between c_1 and c_2 (v_1 and v_2), stabilizing the insulating states for the range of D at $n = \pm n_s$ fillings. For more details, see ref. [23].

Extended Data Fig. 1c, d shows a direct comparison between experimental resistivity and the calculated density of states at the Fermi energy for the $\theta = 1.33°$ TDBG. The experimental results are plotted against displacement field D while the calculation is plotted against onset potential difference between the top-most and bottom-most graphene layer U. The conversion between the experimental parameter D and the calculation parameter U is not straightforward owing to the screening of the electric field by the graphene layers themselves. Thus, converting D to U requires a self-consistent calculation of the screening effect produced by the TDBG band structure, which in turn depends on U. However, in Extended Data Fig. 1, we see a very good match between experimental single-particle insulating states and theoretical gaps at the Fermi energy when we convert D into U with an empirical factor: $U = 0.1\ \text{nm} \times eD$, where e is the electron charge. Besides the single-particle gaps, the calculation also shows regions of high density of states. In particular, in experimental data, there are two lines with higher resistivity than the surroundings: from $(n, D) = (-1, -0.2)$ to $(1, 0.6)$ and from $(n, D) = (-1, 0.2)$ to $(1, -0.6)$. These two lines form a cross and pass through the half-filled insulator as well as the halo features. By comparing with the calculation, we recognize that these experimental features correspond to the regions of high density of states shown in Extended Data Fig. 1d.

Hall effect and ferromagnetism phase boundary

In Extended Data Fig. 2, we show the Hall effect measured in the sample with the most robust half-filled insulating states ($\theta = 1.41°$, $\Delta_{n_s/2} = 4.2$ meV). There is a clear change in Hall resistance behaviour across the halo boundary, which is identifiable in the longitudinal resistance measurement (Extended Data Fig. 2a). Inside the halo, the Hall resistance changes sign across the half-filled insulating state, with a positive value above half-filling and a negative value below half-filling. This Hall measurement demonstrates that the metallic states inside the halo are closely related to the half-filled insulator, probably by a change of Fermi level from half-filling to the inside of the subband (Extended Data Fig. 2c). Given that the half-filled insulator is spin polarized, the metallic states are probably a ferromagnetic metal, which contains two spin-polarized bands that are shifted in energy by the ferromagnetic exchange coupling ((2) in Extended Data Fig. 2c). Outside the halo, the Hall effect follows the expectation of a single-particle moiré band without correlation effects such as in large-angle TBG, indicating that the system recovers to a normal metallic state.

In Extended Data Fig. 2a, we also notice that a three-quarter-filled insulating state appears on the border of the halo. The simplest possible candidate of this state is a spin- and valley-polarized state. However, within a simple mean-field picture, we expect the lowest-lying excitations in this state to be associated with valley-flip rather than spin-flip as the spin exchange coupling is expected to be larger than the valley exchange splitting[23]. This naive picture appears to be inconsistent with the enhancement of the gap by the in-plane magnetic field shown in Fig. 2d in the main text. In addition, the appearance of the quarter-filling state right at the edge of the halo suggests that a mean-field picture may fail to capture the relevant physics. We leave this question regarding the nature of the quarter-filling state to future theoretical works.

Magnetic-field-induced Chern insulator state in the 1.26° sample

In the $\theta = 1.26°$ sample, we observed distinctly different behaviour of the Hall resistance. Under a small perpendicular magnetic field, the Hall resistance is always positive inside the halo (Extended Data Fig. 3b), rather than changing signs across half-filling. The absent sign change of the Hall signal across half-filling may be due to thermal excited carriers of both types as a result of the small insulating gap. It could also be due to the Chern insulator behaviour discussed below. Measuring the Hall resistance with changing magnetic field and density at a fixed displacement field (Extended Data Fig. 3d) reveals a single line of large Hall signal tracing to half-filling with a slope corresponding to $\nu = 4$. At the same time, longitudinal resistance develops a minimum along the same line (Extended Data Fig. 3c). Following the $\nu = 4$ line (black guiding lines in Extended Data Fig. 3c, d), Extended Data Fig. 3e shows the Hall resistance reaches close to a quantized value of $h/4e^2$ when the perpendicular magnetic field $B_\perp > 3$ T.

The fact that a single Hall plateau emerges from half-filling strongly suggests that this is not a normal quantum Hall state. Instead, the data highly resemble the Chern insulator shown in MA-TBG[17]. Indeed, our theory predicts that in TDBG, the first conduction band has a Chern number $C = 2$ in one valley and an opposite Chern number $C = -2$ in the other valley[23]. As shown in the main text, without a perpendicular magnetic field, the half-filled state is spin polarized and valley unpolarized, giving a total Chern number of 0. However, a perpendicular magnetic field couples to the valley degree of freedom through the orbital valley Zeeman effect. When the spin-polarized gap is small such as in the 1.26° device, the valley Zeeman energy can overcome the spontaneous spin-polarized gap and converts the spin-polarized half-filled insulating state into a valley-polarized Chern insulator. Using the valley Zeeman factor from an scanning tunnelling microscopy study and calculation[23], we estimate that valley Zeeman energy surpasses the 0.3 meV gap at a perpendicular field of 0.2 T. This valley-polarized half-filled state fills two moiré bands (of spin up and spin down) in one valley, adding up to a total Chern number of four.

Critical transition behaviours in the $\theta = 1.26°$ sample

In Extended Data Fig. 4c, the dome of the superconducting-like state, similar to MA-TBG, can be seen next to the half-filled insulator. In addition, cutting through a constant density line, a similar dome structure is visible over the displacement field axis (Extended Data Fig. 4b). The dome in the displacement field terminates on the boundary of the halo. It may first appear that the low-resistance state outside the halo boundary resembles a superconductor as well. However, as we discussed in the main text, there is no critical transition outside the halo. In addition, the I–V characteristic outside the halo is very different from that inside the halo. Within the halo, differential resistance shows a critical current that reduces to zero when approaching the halo boundary (Extended Data Fig. 4f). Outside the halo, in contrast, the I–V characteristic does not fit that of a superconductor (Extended Data Fig. 4g). We believe the low resistance outside the halo is purely caused by ballistic transport. Extended Data Fig. 4d shows that the superconducting-like state has a critical perpendicular magnetic field of about 0.1 T.

In a recent study[31], He et al. observed a sudden drop of resistance with a residue resistance of about 1 kΩ, similar to our 1.32° device. They also reported a nonlinear I–V curve, where $\mathrm{d}V/\mathrm{d}I(I)$ gradually increases with current in a parabolic manner up to a factor of two without signs of critical current. This observation is in stark contrast to the data from our superconducting-like sample, where $\mathrm{d}V/\mathrm{d}I$ reaches zero at zero bias and saturates to a finite value above the critical current. In their paper, He et al. explain this nonlinear I–V as an effect from the temperature increase due to bias current heating. While this argument can explain the nonlinear I–V observed in their sample with large residue resistance, we demonstrate that our superconductor-like nonlinear I–V is unlikely caused by heating.

At 2 K, the zero-bias resistivity of the sample is close to zero. If we assume the sample is metallic, we can translate resistivity to thermal conductivity. Taking the upper bound of resistivity to be 50 Ω, it converts to 1 nW K^{-1} in heat conductivity according to the Wiedemann–Franz law at 2 K. From the I–V curve shown in Fig. 3c, we can extract the heating power to be about 13 pW at the critical current of about 300 nA. As a result, the temperature increase is about 13 mK. In contrast, it requires a 5 K temperature increase to bring the sample resistance to the normal value (Fig. 3b). This provides additional confirmation that our observed nonlinear I–V is probably an intrinsic property of the device rather than a heating effect.

Enhancement of the transition temperature with $B_{\parallel}$

If the superconducting-like behaviour in the 1.26° sample is indeed from superconductivity, the parallel field dependence shown below suggests that it might be an exotic spin-polarized superconductor. Here we investigate the behaviour of $\rho(T)$ as a function of $B_{\parallel}$. Extended Data Fig. 5a shows a superconducting dome in the $(n, B_{\parallel})$ plane with a maximum critical parallel magnetic field $B_{\parallel}^{c} \approx 1$ T. The salient experimental feature is the $B_{\parallel}$ dependence of the superconducting state below the critical field $B_{\parallel}^{c}$. Extended Data Fig. 5b shows ρ at optimal density and displacement field $(n_{\mathrm{m}}, D_{\mathrm{m}})$ as a function of T and $B_{\parallel}$. In this optimal superconducting state, ρ vanishes critically as T and $B_{\parallel}$ decreases. We use a phenomenological definition of the critical temperature $T_{50\%}$ defined as the 50% transition point. Interestingly, $T_{50\%}(B_{\parallel})$ follows a non-monotonic behaviour. In particular, $T_{50\%}$ increases as $B_{\parallel}$ increases from 0 to about 0.3 T before it decreases for $B_{\parallel} > 0.3$ T. We also performed I–V characterization at the optimal gate configuration $(n_{\mathrm{m}}, D_{\mathrm{m}})$ as a function of $B_{\parallel}$ and T to obtain T_{BKT} (Extended Data Fig. 4e). Similar to $T_{50\%}$ above, $T_{\mathrm{BKT}}(B_{\parallel})$ also shows a non-monotonic behaviour as shown in Extended Data Fig. 5b (black circles). These sets of evidence suggest that a small $B_{\parallel}$ can strengthen the superconductivity.

The increase of critical temperature with $B_{\parallel}$ suggests that the Cooper pairs responsible for the superconductivity here, if confirmed, are likely to be spin polarized. One possible scenario for such a state is illustrated in Extended Data Fig. 5d, where the Cooper pairs form between Fermi surfaces with the same spin (spin-triplet) and opposite valleys. This model is consistent with our previous discussion of a ferromagnetic metal parent state inside the halo next to the half-filled insulator, where the Fermi surfaces of two different spins have different filling status. In this spin-polarized pairing scheme, a parallel magnetic field enlarges the majority spin Fermi surface, and strengthens the superconductivity, inducing the change in the critical temperature $\Delta T_{\mathrm{c}} \propto B$ (ref. [23]). The eventual destruction of superconductivity at high magnetic fields can result from the following mechanism. Magnetic flux in between layers leads to a momentum shift, which has an opposite sign in the two valleys, thereby bringing the two pairing Fermi surfaces out of alignment. The latter effect is expected to reduce the critical temperature, $\Delta T_{\mathrm{c}} \propto -B^{2}$ (ref. [23]). Alternatively, if the ferromagnetic pairing is caused by spin fluctuations, as suggested in the heavy fermion metals[37–39], a strong parallel magnetic field can suppress the superconductivity by suppressing spin fluctuations[40].

Meanwhile, in the measurement shown in Extended Data Fig. 5c, we notice that the 0 T resistivity goes slightly negative (about −30 Ω) at the lowest temperature for this specific thermal cycle. In general, we find that the four-terminal resistivity we measure in the 1.26° device sometimes shows a small residue (−50–50 Ω) that varies between different thermal cycles. The residual resistance is not present in the measurement shown in Fig. 3b. This residual resistance (when it is present in a specific thermal cycle) is sensitive to the measurement configuration. Extended Data Fig. 6 shows two different four-terminal resistance measurement configurations. Between the two configurations (blue curve and red curve), we essentially flip the direction of the current. As we use a low-frequency a.c. (about 17 Hz) for the current source and a lock-in amplifier for the voltage probe, in an ideal condition, we expect to obtain the same signal with opposite polarity. However, we find that the measured signals deviate from this expectation when there is residual resistance at low temperature. Specifically, when flowing current from top to bottom (blue curve in Extended Data Fig. 6), four-terminal resistance is positive, and a positive residual resistance of about 50 Ω remains at the lowest temperature, 2 K. However, when the current is flowing from the bottom to the top (red curve), the four-terminal resistance at higher temperature is negative as expected, but the residual resistance at lower temperature is still positive and is nearly identical to that in the blue curve (see inset).

We believe that the observed anomaly in the residual resistance originates from bias-induced gating in combination with thermoelectric voltages present in our cryostat wiring. Owing to the temperature gradient in the cryostat, a d.c. thermoelectric voltage is always present between different pairs of wires. This d.c. voltage is simply added to the voltage probes on top of the a.c. voltage induced by bias current. In such cases, the a.c. bias voltage on the sample (about half of the bias voltage on the source lead, as the drain is grounded) can modulate the d.c. thermoelectric voltage through a bias-induced gating effect (see ref. [41] for a similar effect observed in a drag experiment), resulting in a voltage signal synchronizing with the applied a.c. bias current. We can eliminate this bias-induced a.c. gating effect of thermoelectric voltage by subtracting the blue curve from the red curve, where we obtain a near-zero residual resistance within the noise level. During the thermal cycle in obtaining the data in Extended Data Fig. 5c, we did not collect the data in two different configurations, and thus such correction was not possible. Note that this a.c. gating effect becomes appreciable only when the device resistivity is really small (<50 Ω).

Landau fan diagram

In the 1.33° device under a perpendicular magnetic field, clear fans can be identified coming from the CNP, full fillings $\pm n_{\mathrm{s}}$ (about $\pm 4.1 \times 10^{12}$ cm^{-2})

as well as half-filling $n_s/2$ on the electron-doped side (Extended Data Fig. 7a). The Landau fan from the CNP shows a well-developed quantum Hall sequence with fourfold degeneracy on the valence band side under low magnetic fields, and subsequently develops the full degeneracy-lifted quantum Hall states under higher magnetic fields. The Landau fans on the conduction band side ($n > 0$) are highly unusual. The fan from the CNP shows a sequence of only odd filling fractions $v = 3$ and 5. Although the fan coming from the correlated insulating state at half-filling shows a degeneracy of two, consistent with the picture of a spin-polarized half-filled band, the sequence is also of odd numbers $v = 3$, 5 and 7. The odd-integer sequenced quantum Hall effect from the half-filling gap may relate to a Berry phase effect on the Landau level, similar to the quantum Hall effect in monolayer graphene. Theory[23] predicts that the isolated flat conduction band (c_1) in TDBG carries non-trivial Berry curvature and a non-zero Chern number. It is then possible that the Berry curvature accumulates to a π or 3π phase on certain Landau orbits, resulting in a quantum Hall filling fraction sequence of $2(N-1/2)$ or $2(N-3/2)$, with integer N.

There is also a single quantum Hall state $v = 3$ projected down to the quarter-filled conduction band. This state could be a magnetic-field-induced Chern insulator, similar to the $v = 4$ Chern insulator in 1.26° device discussed above. As this fan projects down to quarter-filling, it is likely that both spin and valley are polarized. We note that the valley Chern number corresponding to this specific state is $C = 3$, although $C = 2$ is generally expected[23]. Further study is required to clarify these experimental observations. Above a perpendicular magnetic field of 7 T, the half-filled insulator disappears, presumably due to the orbital effect of the perpendicular magnetic field.

The fan diagram of the 1.26° device is also intriguing. Besides the Landau fans, we point out there is an obvious oscillation of longitudinal resistance around $n = -n_s/2$ that does not sensitively depend on doping (horizontal features marked in Extended Data Fig. 7b). These oscillations are not from quantum Hall states, but from Hofstadter's butterfly. They are a result of the interplay between two different periodicities in the system: moiré superlattice and magnetic length. When one moiré superlattice contains $1/N$ magnetic flux (N is an integer), the two length scales become commensurate and produces a minimum in resistivity. These features are indicative of a highly uniform twisting angle distribution in the sample. We use these features to determine the twist angle in this device.

Effective mass of the $\theta = 1.33°$ and $\theta = 1.26°$ samples

We calculate the effective cyclotron mass from temperature-dependent magnetoresistance (Shubnikov–de Haas (SdH)) oscillations. The cyclotron mass is a measure of the density of states and thus directly related to the Landau-level separation (cyclotron gap) under a given magnetic field. As temperature increases, the SdH oscillation amplitude is reduced following $\Delta R \propto \chi/\sinh(\chi)$, where $\chi = \frac{2\pi^2 k m^*}{\hbar e}\frac{T}{B}$.

For the $\theta = 1.33°$ sample, we measured SdH oscillations at all densities between filling factor $n/n_s = -1$ and 1, at $T = 0.3$, 2, 3, 4, 6, 9 and 14 K (example: Extended Data Fig. 8a–c). We then extracted the oscillation amplitudes and plotted them as a function of T/B. Fitting $\Delta R(T/B)$ with the above formula with m^* being the only fitting parameter, we obtained the effective cyclotron mass shown in the main text. Similarly, we measured SdH oscillations for the $\theta = 1.26°$ sample and extracted an effective mass $m^* = 0.23m_e$, as shown in Extended Data Fig. 8e, f.

Device fabrication and characterization

All the devices presented in this study were prepared using the dry transfer method[42], using stamps consisting of polypropylene carbonate film and polydimethylsiloxane. Half of a bilayer graphene flake was torn and picked up by a stack of graphite/hBN on the transfer stamp. Then the remaining bilayer graphene flake on the substrate was rotated by the desired angle and picked up. The stacks were deposited on a 300-nm SiO_2/Si substrate after picking up the rest of hBN and graphite layers. Part of the bilayer graphene flakes was extended outside the hBN area onto polypropylene carbonate to prevent the graphene from freely rotating on the hBN. The resulting stacks were fabricated into 1–2-µm-wide devices to ensure a uniform twist angle in the relatively narrow channel. The temperature of the stack was always kept below 180 °C during the stacking and fabrication processes.

We measured a total of seven devices with different twist angles in this study. All samples were encapsulated by the hBN layers. In the 1.32°, 1.33°, 1.41°, 1.53° and 2.00° devices, both the top and bottom gates were made from graphite. The 1.26° device utilized a graphite top gate and a heavily doped silicon bottom gate below the hBN substrate and 300-nm SiO_2 dielectric. The 1.48° device used a silicon bottom gate and a metal top gate. Most of the devices were fabricated into Hall bars with the exception of the 1.26° and 1.41° samples, which were fabricated into a Van der Pauw geometry. The gate configurations and device images are shown in Extended Data Fig. 9.

The resistivity presented here was measured at 17.7 Hz using the standard lock-in technique, with a 0.5–1.0 mV voltage bias and a current-limiting-resistor of 100 kΩ connected in series with the sample, which limits the current in the sample to an upper bound of 5–10 nA. This bias scheme is to ensure neither the voltage nor the current becomes too large when sweeping across states with drastically different resistance (insulators or superconducting-like states). The four-terminal voltage and the source–drain current are measured simultaneously with two lock-in amplifiers to obtain the four-terminal resistance. Resistivity is then obtained by multiplying the resistance by a geometric factor (about 4.5 for Van der Pauw devices).

Figure 1f and Extended Data Fig. 9a, c–g show large-scale gate scans of the longitudinal resistivity in all samples. These samples show insulating states at full filling under the zero displacement field and at the CNP under a large displacement field. Moreover, in particular, the 1.32°, 1.48°, 1.53° and 2.00° devices show a CNP gap under the zero displacement field, which closes and reopens with increasing displacement field. For the 1.32° device, we note that the gate scan diagram shows a wide insulating region at full filling and double peaks at half-filling, suggesting that there is additional moiré periodicity in the channel other than 1.32°. The CNP gap in the 1.32° device under zero displacement field probably originated from the larger twist angle region of the sample, as none of the neighbouring angle devices (1.26, 1.33, 1.41°) has a gap at the CNP under zero displacement field. We remark that theory[23] predicts a gap at the CNP opens up even at zero displacement field when $\theta > 1.5°$, qualitatively agreeing with our experimental observation.

The twist angles are estimated from two independent methods. For the first method, we measure the gate voltages of full-filling gaps and convert these voltages to full-filling density n_s using the gate capacitance. The gate capacitance used for this conversion is calibrated by Landau fan diagrams. We then calculate the twist angle from the fact that full filling corresponds to four electrons per moiré unit cell so the moiré unit cell area $A = 4/n_s$. The main source of errors for this method is in locating the exact position of full filling in the gate voltage. We also use the radiating Landau fan coming from full filling to help locate its exact position. Typically, we can identify the position of full filling to an accuracy of $\delta n_s \approx 10^{11}\ \mathrm{cm}^{-2}$, corresponding to a twist angle error of ±0.02°. The second method exploits Hofstadter's butterfly features under magnetic fields. Here we find carrier-density-independent oscillations of the longitudinal resistance R_{xx} under perpendicular magnetic fields, with the minimum of R_{xx} appearing when the magnetic flux in a moiré unit cell is a fraction of the flux quantum, $BA = \phi_0/N$ (Extended Data Fig. 7b), where B is magnetic field, ϕ_0 is the flux quantum and N is an integer. These features are used to calculate the twist angle with even better accuracy (±0.01°).

In Extended Data Fig. 9a, we observe negative resistivity in a part of the 2D gate scan. These anomalous signals can be attributed to non-transparent contacts in these gate regions. Comparing with the 2D map of the two-terminal resistance measured in this device (Extended Data Fig. 9b), we find that the gate regions where negative resistivity were observed in Extended Data Fig. 9a in general correspond to the gate regions where a large contact resistance is demonstrated by the two-terminal measurement. This strong correlation between the apparent 'negative' four-terminal resistivity and high contact resistance suggests that the negative four-terminal resistivity originates from inefficient contact equilibration in these gate regimes. Indeed, the contact transparency can be hindered by the unintended p–n junction formation near the metal contacts when the applied gate voltages conspire with the work function mismatch between graphene and metal to cause an accumulation of the opposite polarity of charges near the contact and in the channel. However, near the half-filled insulator region, where most of our research is focused, the two-terminal resistance stays less than 10 kΩ, demonstrating excellent contact transparency. Thus, we conclude that the contact anomaly can be excluded as a possible cause of our experimental observation near the correlated insulator regime. We also notice that the correlated insulating state in the positive displacement field side ($V_{TG} < 0$ and $V_{BG} > 0$) shows a much weaker signature than the negative displacement field side. The absence of a clear signature of correlated insulator on the opposite side of the displacement field is also likely due to the inefficient contact equilibration in these particular gate configurations, as shown in Extended Data Fig. 9b.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

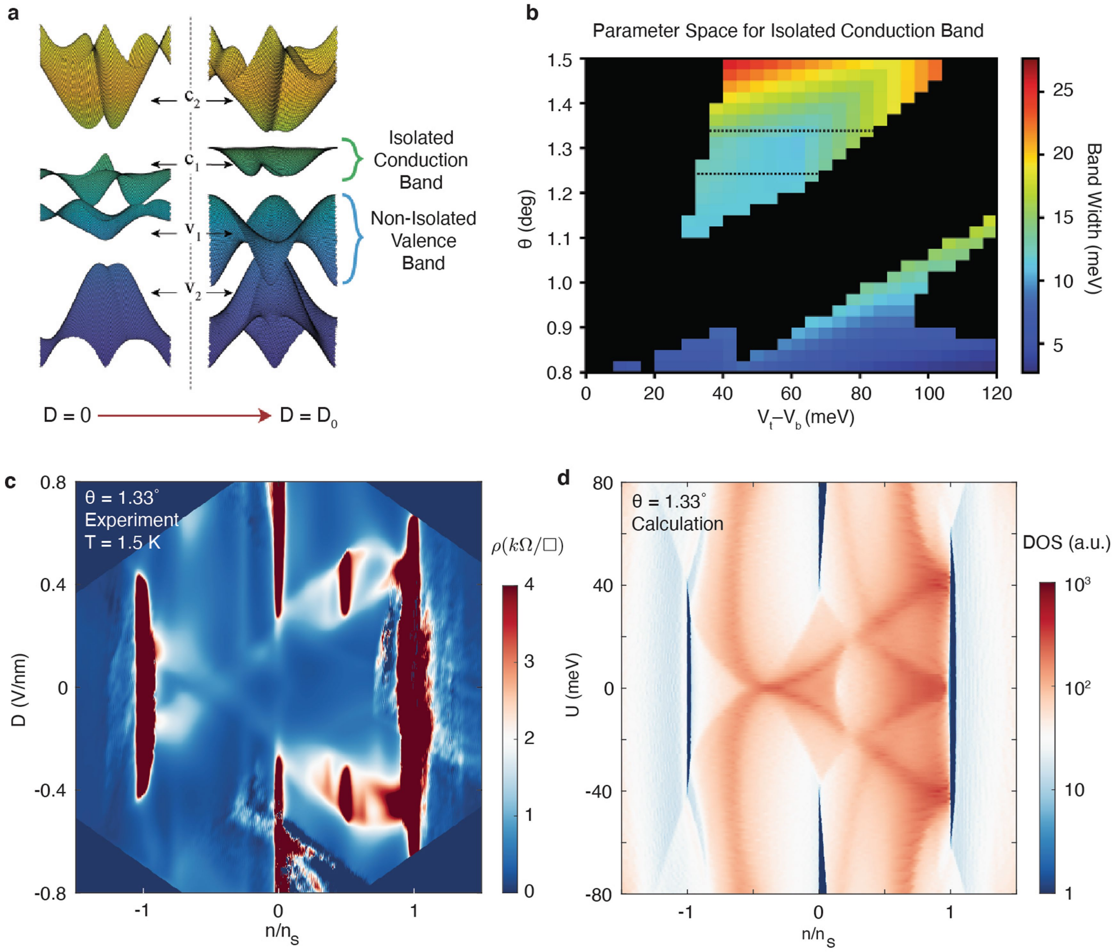
35. Jung, J. & MacDonald, A. H. Accurate tight-binding models for the π bands of bilayer graphene. *Phys. Rev. B* **89**, 035405 (2014).
36. Moon, P. & Koshino, M. Optical absorption in twisted bilayer graphene. *Phys. Rev. B* **87**, 205404 (2013).
37. Saxena, S. S. et al. Superconductivity on the border of itinerant-electron ferromagnetism in UGe_2. *Nature* **406**, 587–592 (2000).
38. Aoki, D. et al. Coexistence of superconductivity and ferromagnetism in URhGe. *Nature* **413**, 613–616 (2001).
39. Huy, N. T. et al. Superconductivity on the border of weak itinerant ferromagnetism in UCoGe. *Phys. Rev. Lett.* **99**, 067006 (2007).
40. Julian, S. Viewpoint: pairing with spin fluctuations. *Physics* **5**, 17 (2012).
41. Liu, X. et al. Frictional magneto-Coulomb drag in graphene double-layer heterostructures. *Phys. Rev. Lett.* **119**, 056802 (2017).
42. Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).

Acknowledgements The major experimental work is supported by DOE (DE-SC0012260). P.K. acknowledges support from the DoD Vannevar Bush Faculty Fellowship N00014-18-1-2877. Z.H. is supported by ARO MURI (W911NF-14-1-0247). A.V., J.Y.L. and E.K. were supported by a Simons Investigator Fellowship. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by the MEXT, Japan, A3 Foresight by JSPS and the CREST (JPMJCR15F3), JST. A portion of this work was performed at the National High Magnetic Field Laboratory, which is supported by the National Science Foundation Cooperative Agreement Number DMR-1157490* and the State of Florida. Nanofabrication was performed at the Center for Nanoscale Systems at Harvard, supported in part by an NSF NNIN award ECS-00335765. We thank S. Fang, S. Carr, Y. Xie, E. Kaxiras, B. I. Halperin, A. F. Young, J. Waissman and A. Zimmerman for helpful discussion.

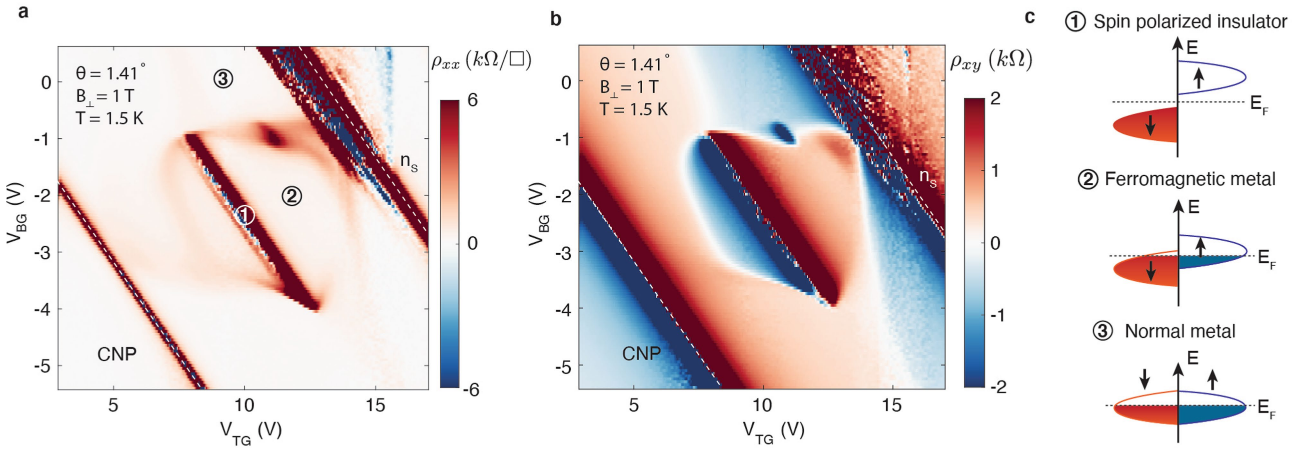
Author contributions X.L. and P.K. conceived the experiment. X.L., Z.H., Y.R., H.Y. and D.H.N. fabricated the samples. X.L. and Z.H. performed the measurements and analysed the data. E.K., J.Y.L. and A.V. conducted the theoretical analysis. X.L., Z.H. and P.K. wrote the paper with input from Y.R., H.Y., E.K., J.Y.L. and A.V. K.W. and T.T. supplied hBN crystals.

Competing interests The authors declare no competing interests.

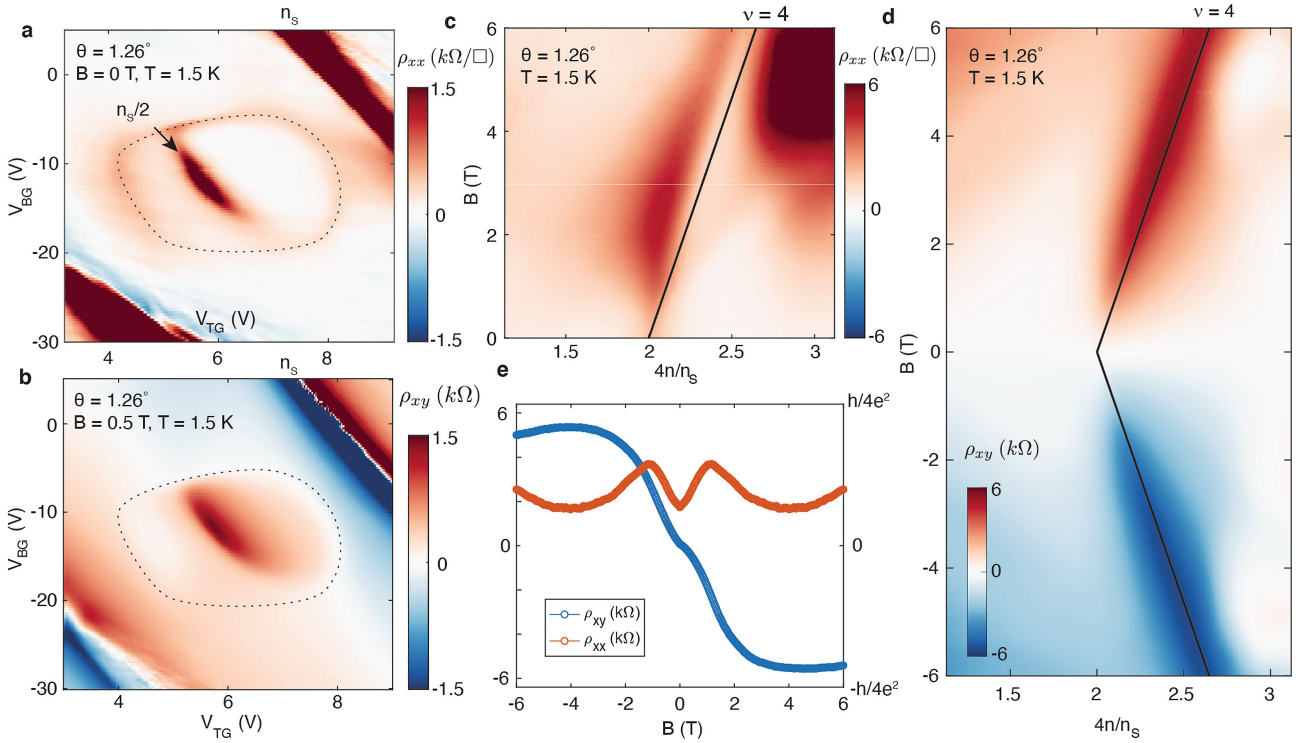
Additional information
Correspondence and requests for materials should be addressed to X.L. or P.K.
Peer review information *Nature* thanks Hu-Jong Lee, Ming-Hao Liu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



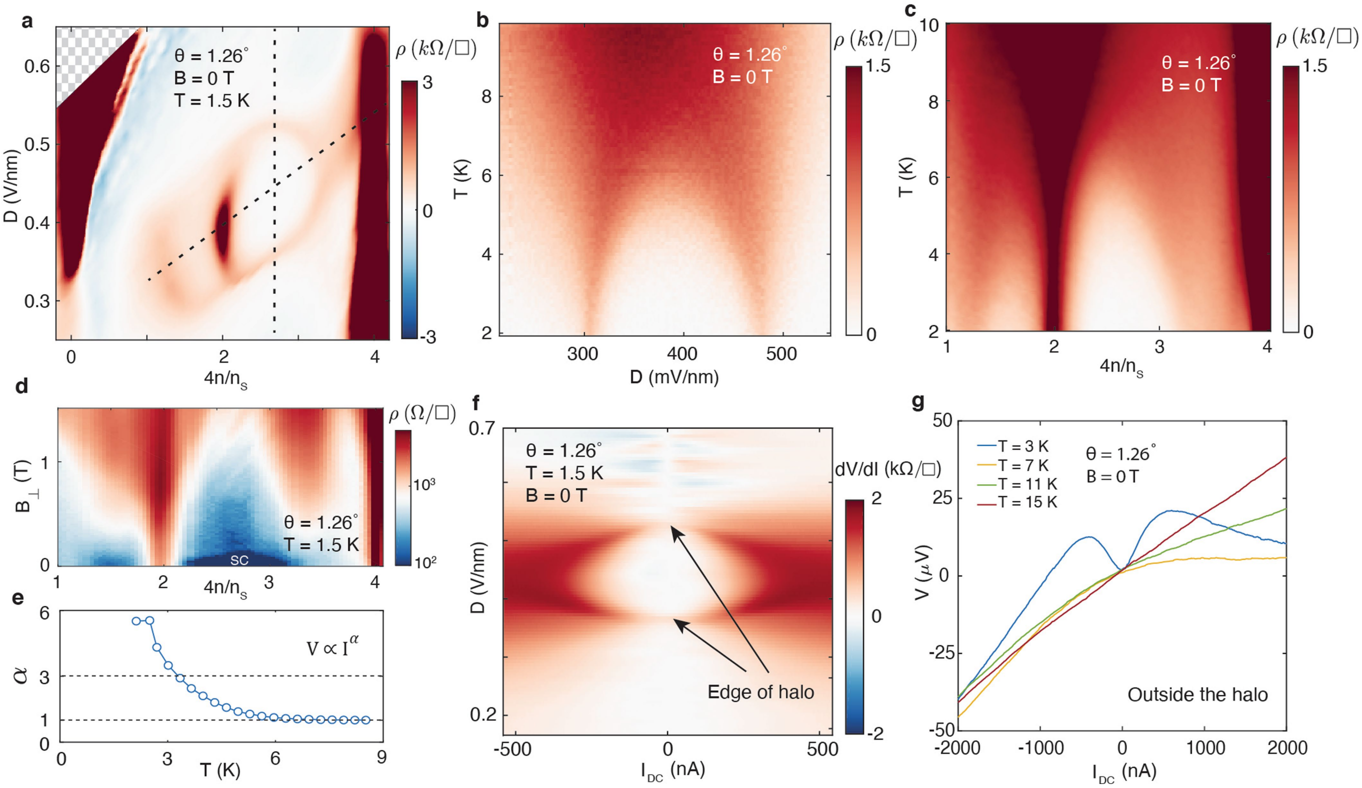
Extended Data Fig. 1 | Theoretical band structure of TDBG. **a**, Calculated band structure of TDBG at zero displacement field and optimal displacement field D_0 for the isolated flat band. **b**, Calculated parameter space for isolated conduction band (x axis is onsite potential difference $U = V_t - V_b$ between the top and bottom graphene layer, y axis is twist angle). Colour represents the bandwidth of the first conduction band c_1 (meV). In the coloured parameter space, c_1 is isolated from the second conduction band and the first valence bands. The two dotted lines represent cuts at $\theta = 1.26$ and $\theta = 1.33°$. **c**, Resistivity as a function of filling fraction and displacement field in the $\theta = 1.33°$ sample. A cross-like feature of high resistivity is formed along two lines from $(n, D) = (1, -0.2)$ to $(1, 0.6)$ and $(-1, 0.2)$ to $(1, -0.6)$, passing through the half-filled insulating states. **d**, Density of states at the Fermi energy calculated by the continuum model. The single-particle insulators ($n/n_s = 0, \pm 1$) in experiment match well with the gaps shown in the calculation and the van Hove singularity captures the cross-like pattern in experiment.



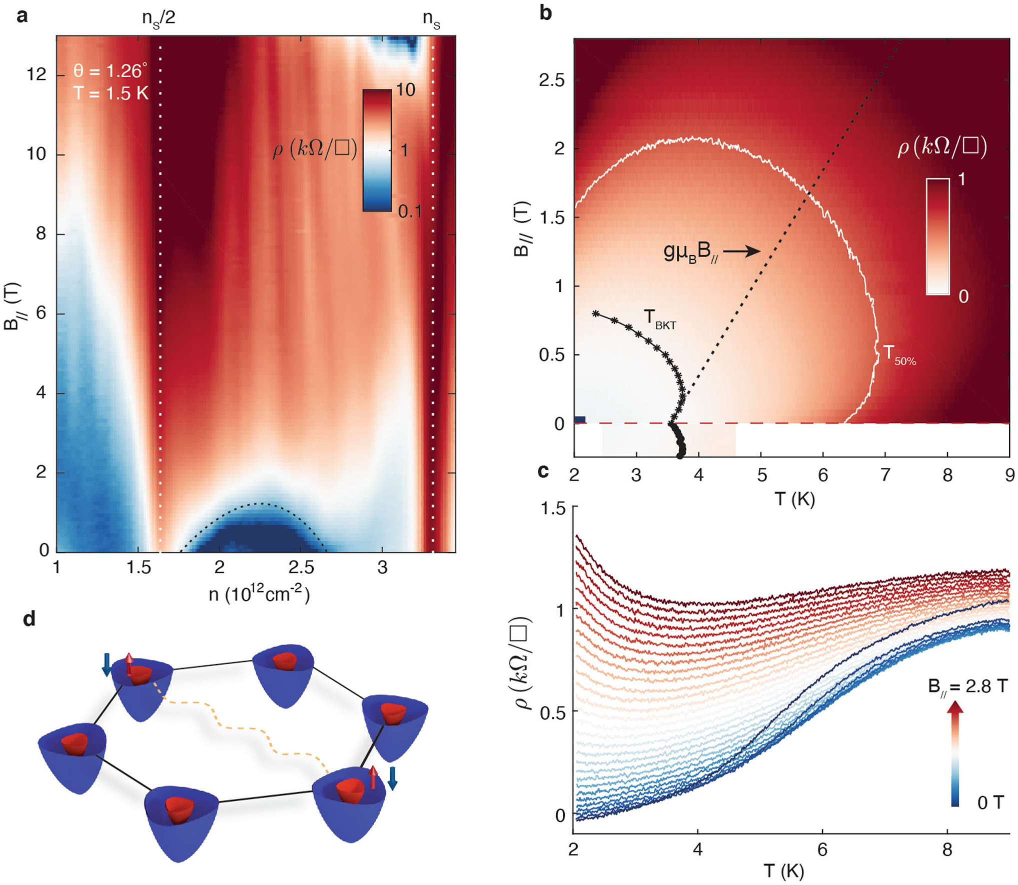
Extended Data Fig. 2 | Hall effect in a device with robust half-filled insulators. **a**, **b**, Longitudinal resistivity (**a**) and Hall resistance (**b**) of the $\theta = 1.41°$ device around half-filling at $T = 1.5$ K and under perpendicular magnetic field $B_\perp = 1$ T. Data are symmetrized between positive and negative fields to eliminate mixing. The halo structure is apparent around the half-filled insulator and a three-quarter-filled insulating state resides on the border of the halo. The Hall resistance changes sign across the half-filled insulator inside the halo. **c**, Illustration of electron orders for different regimes. The left half (right half) of the cartoon represents the band of spin down (up) electrons. For half-filling, only one species of spin is filled. Inside the halo, one spin species is populated more than the other. Outside the halo, both spins are equally populated.



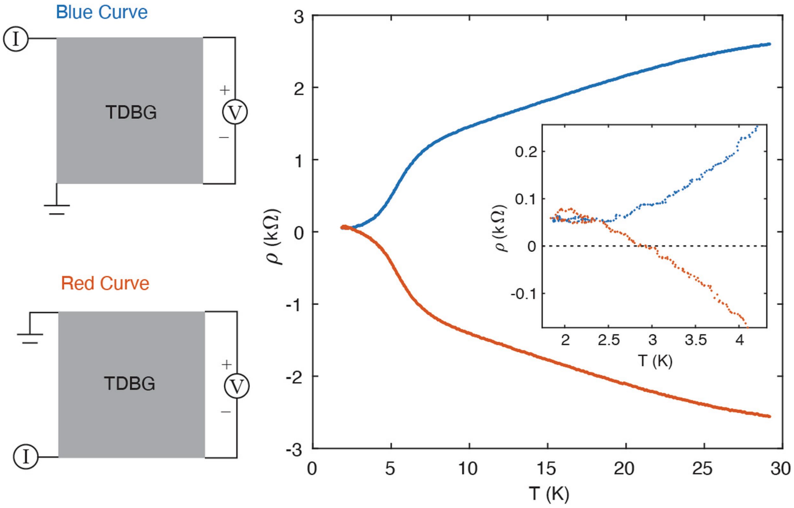
Extended Data Fig. 3 | Field-induced Chern insulator in the $\theta = 1.26°$ device. **a**, **b**, Longitudinal resistivity at $B = 0$ (**a**) and Hall resistance at $B_{\perp} = 0.5$ T (**b**) in the $\theta = 1.26°$ sample at $T = 1.5$ K. The Hall resistance here is symmetrized with both directions of the magnetic field. **c**, **d**, Fan diagram of longitudinal resistivity (**c**) and Hall resistance (**d**) at $T = 1.5$ K at a constant displacement field. The black line marks the expected position for $\nu = 4$ Chern insulator state originating from half-filling. **e**, Longitudinal resistivity and Hall resistance along the black line shown in **c** and **d**.



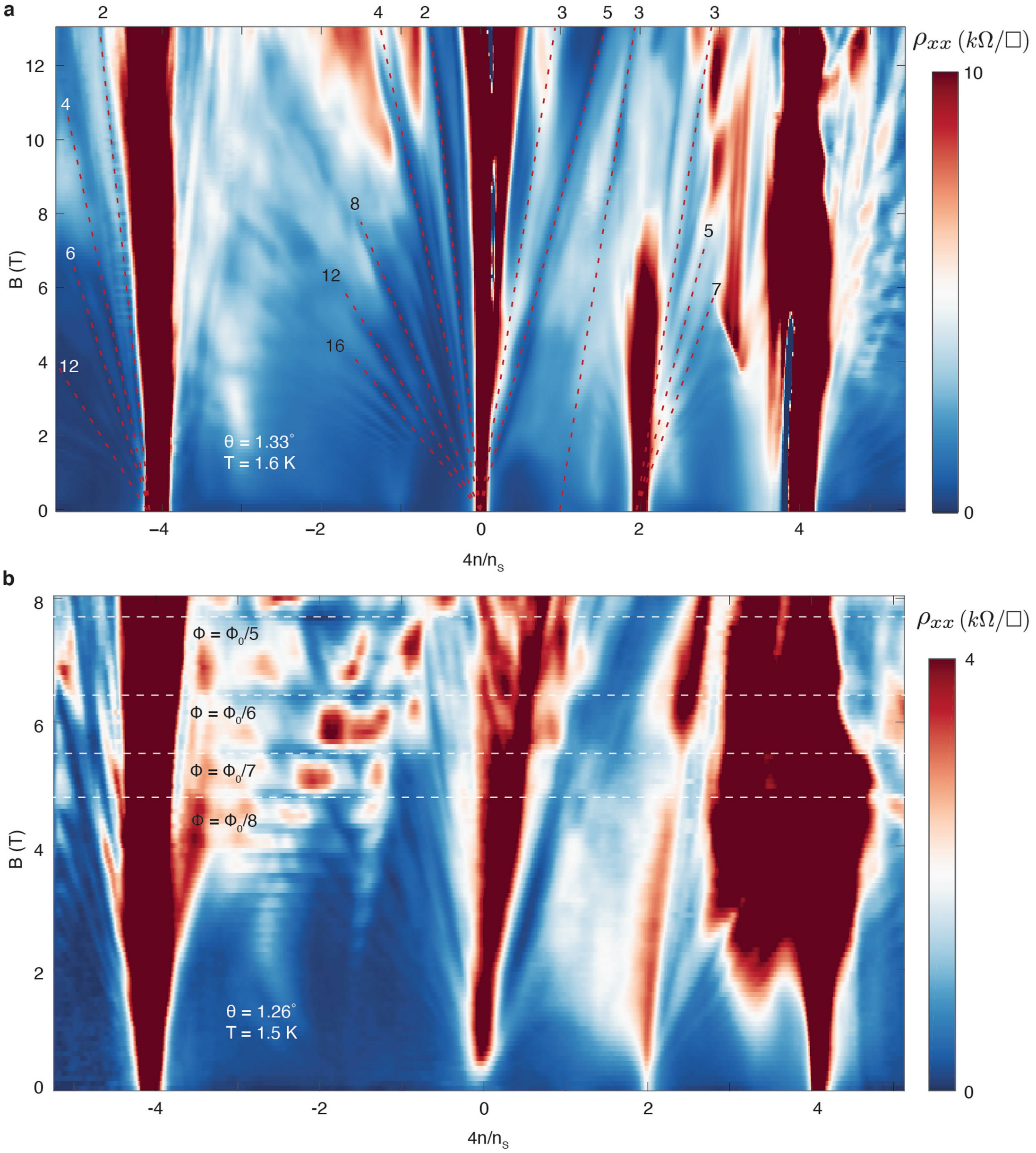
Extended Data Fig. 4 | Critical behaviours in the $\theta = 1.26°$ device. **a**, Resistivity in 1.26° device plotted against filling factor and displacement field. **b**, Resistivity as a function of displacement field and temperature along the constant density line shown in **a**. **c**, Resistivity as a function of filling and temperature along the tilted line in **a**. The dome of the low resistance state can be seen next to the half-filled insulator. **d**, Resistivity on a log scale as a function of filling and perpendicular magnetic field. **e**, The power α in $V \propto I^{\alpha}$ as a function of temperature from fitting the top left inset of Fig. 3c. $\alpha = 3$ is defined as the BKT transition temperature. **f**, Differential resistance as a function of current and displacement field along the constant density line shown in **a**. **g**, I–V curves outside the halo.



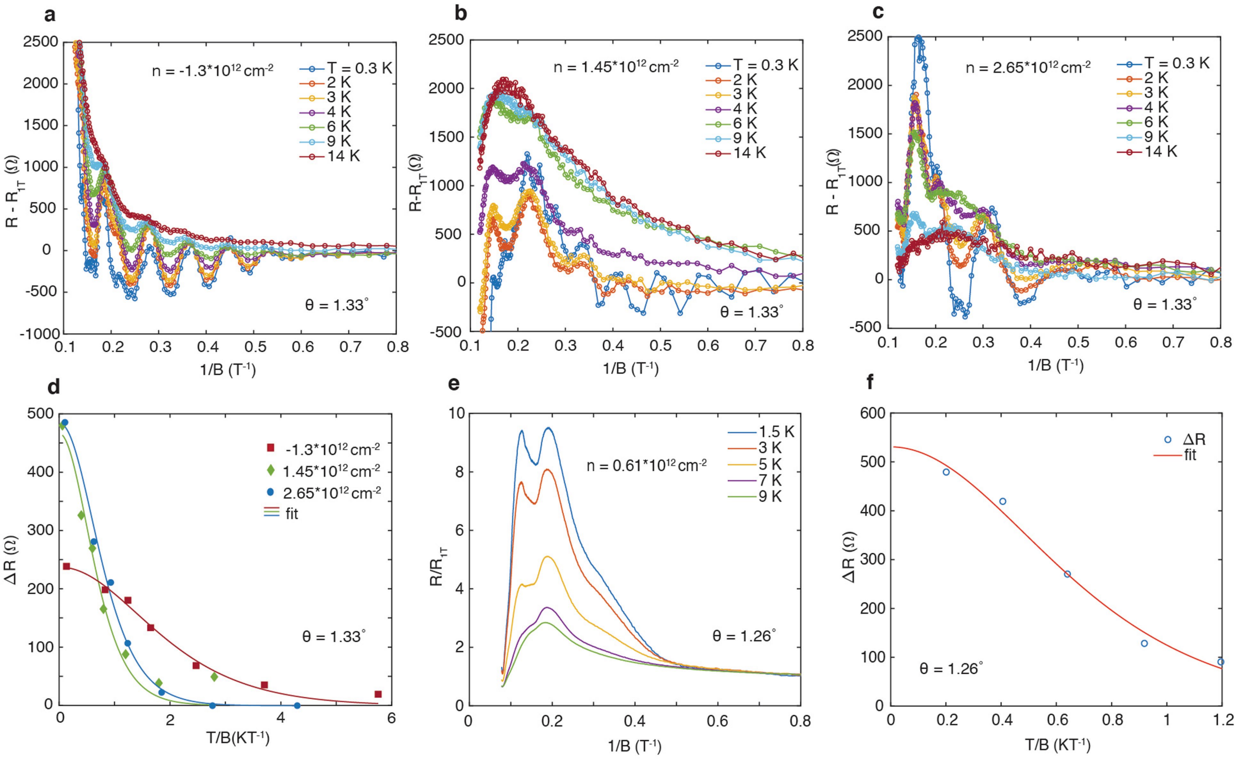
Extended Data Fig. 5 | Enhancement of the critical temperature under in-plane magnetic field. a, Resistivity as a function of in-plane magnetic field across the half-filled insulator and superconducting-like state in the 1.26° device. **b**, Resistivity as a function of temperature and in-plane magnetic field at optimal doping and displacement field. T_{BKT} denotes the BKT temperature extracted from nonlinear IV measurements. $T_{50\%}$ marks the temperature where resistance is half of the normal resistance. **c**, Line traces at different in-plane magnetic fields. **d**, Illustration of pairing in spin-polarized superconductor. The blue (red) surface represents the spin down (up) electron band. The two bands are filled differently due to the parent ferromagnetic metallic state. The hexagon represents the Brillouin zone of graphene lattice. Pairing thus happens between Fermi surfaces of the same spin and opposite valleys.



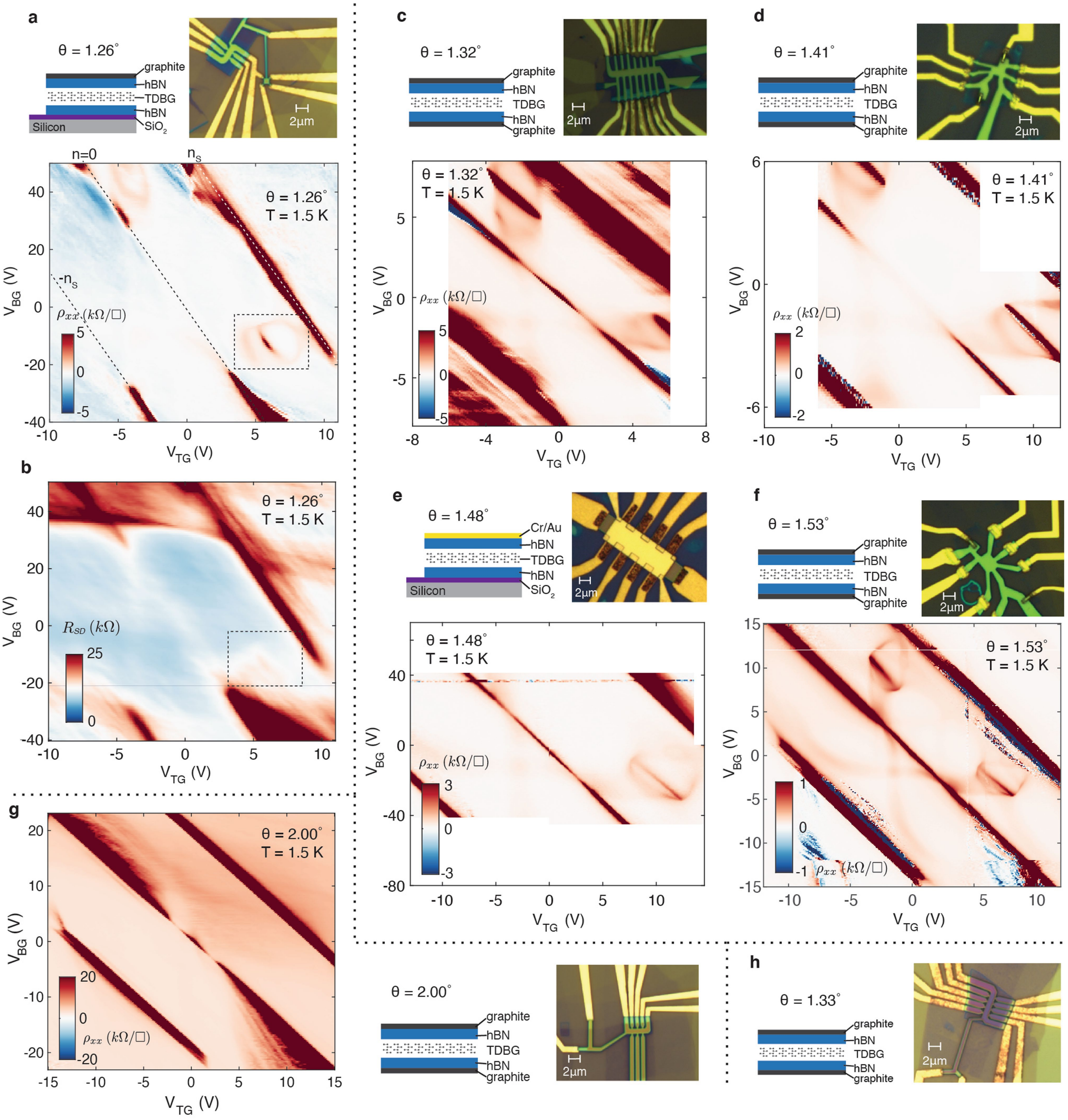
Extended Data Fig. 6 | Origin of a small residual resistivity at $T \ll T_c$. $R(T)$ curve measured in the superconducting regime of the 1.26° device in two measurement configurations. The voltage probes are kept the same between the two configurations while the source and the drain contacts are switched.



Extended Data Fig. 7 | Landau fan diagram as a function of filling fraction and perpendicular magnetic field. a, The 1.33° device. The numbers next to the guiding lines indicate Landau-level filling factors. **b**, The 1.26° device. Horizontal lines highlight the Hofstadter's butterfly features that occur when a simple fraction of the flux quantum ϕ_0/N (N is an integer) penetrates through a moiré unit cell.



Extended Data Fig. 8 | Effective mass calculation for the 1.33° and 1.26° devices. **a**–**c**, Temperature-dependent SdH oscillations in the $\theta = 1.33°$ device at a few representative density points: $n = -1.3 \times 10^{12}$ cm^{-2} (**a**), 1.45×10^{12} cm^{-2} (**b**) and 2.65×10^{12} cm^{-2} (**c**). **d**, Extracted oscillation amplitudes as a function of T/B for the density configuration shown in **a**–**c** and corresponding fitting curves. **e**, Temperature-dependent SdH oscillations in the $\theta = 1.26°$ device at $n = 0.61 \times 10^{12}$ cm^{-2}, which is above half-filling and inside the halo. **f**, Extracted oscillation amplitudes as a function of T/B in the $\theta = 1.26°$ device.



Extended Data Fig. 9 | Device characterization. **a**, **c**–**g**, Device structure, optical image and four-terminal resistivity map of each device: 1.26° (**a**), 1.32° (**c**), 1.41° (**d**), 1.48° (**e**), 1.53° (**f**) and 2.00° (**g**). For the 1.26° device, the active device is the four-terminal Van der Pauw sample. The structure of each device is depicted by the cross-section illustration on the left of the optical image. **b**, Two-terminal resistance measured in the 1.26° device in the same gate voltage range presented in **a**. Dashed square marks the active area studied. **h**, Structure and optical image of the 1.33° device.

Extended Data Table 1 | Summary of all TDBG devices studied

θ	1.26°	1.32°	1.33°	1.41°	1.48°	1.53°	2.00°
w (meV)	12 (11)	13	13 (10)	17	21	24	71
$\Delta_{n_s/2}$ (meV)	0.30	2.8	3.0	4.2	0.54	0.72	N.A.
ρ_{min} (kΩ/□)	<0.01	1.04	0.82	0.21	0.28	0.14	N.A.

Minimum bandwidth w at the optimal displacement field obtained from continuum model calculation (experimentally estimated bandwidth is shown in the bracket), half-filled gap $\Delta_{n_s/2}$ and resistivity well below the critical transition inside the halo $\rho_{\min}$ for devices with different twist angles. There is no sign of any correlated state in the 2.00° device (N.A.). $\Delta_{n_s/2}$ shows a general trend of diminishing away from the optimal angle 1.41°, although the disorder might cause some device-to-device variation.

Large-scale integration of artificial atoms in hybrid photonic circuits

https://doi.org/10.1038/s41586-020-2441-3

Received: 13 October 2019

Accepted: 2 April 2020

Noel H. Wan[1,4✉], Tsung-Ju Lu[1,4✉], Kevin C. Chen[1], Michael P. Walsh[1], Matthew E. Trusheim[1], Lorenzo De Santis[1], Eric A. Bersin[1], Isaac B. Harris[1], Sara L. Mouradian[1,3], Ian R. Christen[1], Edward S. Bielejec[2] & Dirk Englund[1✉]

A central challenge in developing quantum computers and long-range quantum networks is the distribution of entanglement across many individually controllable qubits[1]. Colour centres in diamond have emerged as leading solid-state 'artificial atom' qubits[2,3] because they enable on-demand remote entanglement[4], coherent control of over ten ancillae qubits with minute-long coherence times[5] and memory-enhanced quantum communication[6]. A critical next step is to integrate large numbers of artificial atoms with photonic architectures to enable large-scale quantum information processing systems. So far, these efforts have been stymied by qubit inhomogeneities, low device yield and complex device requirements. Here we introduce a process for the high-yield heterogeneous integration of 'quantum microchiplets'—diamond waveguide arrays containing highly coherent colour centres—on a photonic integrated circuit (PIC). We use this process to realize a 128-channel, defect-free array of germanium-vacancy and silicon-vacancy colour centres in an aluminium nitride PIC. Photoluminescence spectroscopy reveals long-term, stable and narrow average optical linewidths of 54 megahertz (146 megahertz) for germanium-vacancy (silicon-vacancy) emitters, close to the lifetime-limited linewidth of 32 megahertz (93 megahertz). We show that inhomogeneities of individual colour centre optical transitions can be compensated in situ by integrated tuning over 50 gigahertz without linewidth degradation. The ability to assemble large numbers of nearly indistinguishable and tunable artificial atoms into phase-stable PICs marks a key step towards multiplexed quantum repeaters[7,8] and general-purpose quantum processors[9–12].

Artificial atom qubits in diamond combine spin clusters with minute-scale coherence times[5] and efficient spin–photon interfaces[2], making them attractive for processing and distributing quantum information[1,3]. In particular, proposed quantum repeaters for long-range, high-speed quantum networks will require hundreds or more memory qubits[1,7], whereas error-corrected quantum computing may require millions or more[9–12]. However, a critical barrier to large-scale quantum information processing is the low device yield of functional qubit systems. Furthermore, although individual spin–photon interfaces can now achieve excellent performance, the lack of active chip-integrated photonic components and wafer-scale, single-crystal diamond currently limit the scalability of monolithic diamond quantum-information-processing architectures. A promising method to alleviate these constraints is heterogeneous integration, which is increasingly used in advanced microelectronics to assemble separately fabricated subcomponents into a single, multifunctional chip. Such hybrid fabrication has also recently been used to integrate photonic integrated circuits (PICs) with quantum modules, including quantum dot single-photon sources[13–15], superconducting nanowire single-photon detectors (SNSPDs)[16] and nitrogen-vacancy (NV) centre diamond waveguides[17]. However, these demonstrations assembled components one by one, which presents a formidable scaling challenge, and they did not provide for spectral alignment of artificial atoms. Here we introduce a 'quantum microchiplet' (QMC) framework that greatly improves the yield and accuracy of heterogeneously integrated nanoscopic devices. Specifically, this assembly process enables the construction of a 128-channel photonic integrated artificial atom chip containing diamond quantum emitters with high coupling efficiencies, optical coherences near the lifetime limit and tunable optical frequencies to compensate for spectral inhomogeneities on chip.

Figure 1 illustrates the heterogeneous integration process. The multichip module consists of a waveguide layer in single-crystal aluminium nitride (AlN) for low-loss photonics[18], microchiplet sockets to optically interface with separately fabricated diamond QMCs and electrical layers for controlling colour centre transitions. This PIC platform is compatible with additional components, such as on-chip electro-optic modulators[19] and SNSPDs[16,20], for photon switching and photon detection in a quantum photonic chip.

---

[1]Research Laboratory of Electronics, MIT, Cambridge, MA, USA. [2]Sandia National Laboratories, Albuquerque, NM, USA. [3]Present address: University of California Berkeley, Berkeley, CA, USA. [4]These authors contributed equally: Noel H. Wan, Tsung-Ju Lu. ✉e-mail: noelwan@mit.edu; tsungjul@mit.edu; englund@mit.edu

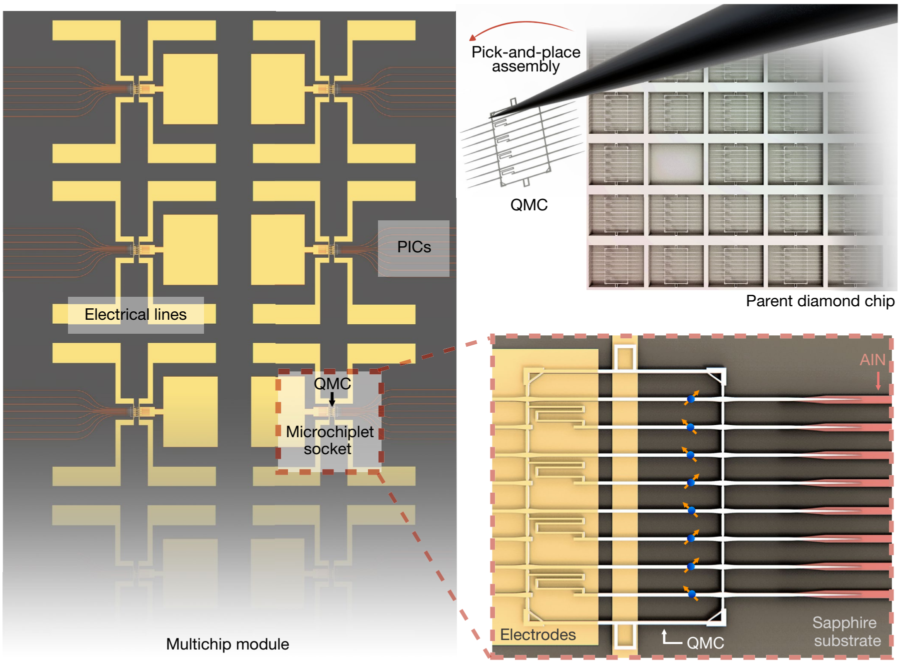


Fig. 1 | Scalable integration of artificial atoms with photonics. The separate fabrication of subcomponents before their final assembly maximizes the yield, size and performance of the hybrid emitter–photonics chip. A pick-and-place method transfers pre-screened QMCs from their parent diamond chip into a socket containing efficient photonic interfaces, as well as electrical wires for controlling colour centres.

The large optical transparency of the QMC and PIC materials make them compatible with a variety of quantum emitters. Here we consider the negatively charged germanium-vacancy (GeV) and silicon-vacancy (SiV) centres in diamond with zero-phonon line transitions at 602 nm and 737 nm, respectively, because of their stable optical and spin properties[21–25]. The process begins with focused ion beam (FIB) implantation of Ge^+ and Si^+ into a 1-µm-pitch square array in a single-crystal diamond substrate, followed by high-temperature annealing (Methods). This process generates spots of tightly localized GeV centres (depth of about 74 nm, vertical straggle of about 12 nm and lateral full-width at half-maximum (FWHM) distribution of about 40 nm) and SiV centres (about 113 nm, about 19 nm and about 50 nm, respectively), which we then located and mapped relative to prefabricated alignment markers by photoluminescence microscopy. We fabricated the QMCs over the emitter arrays using a combination of electron-beam lithography (EBL) and quasi-isotropic etching[26,27]. Figure 2a shows a scanning electron

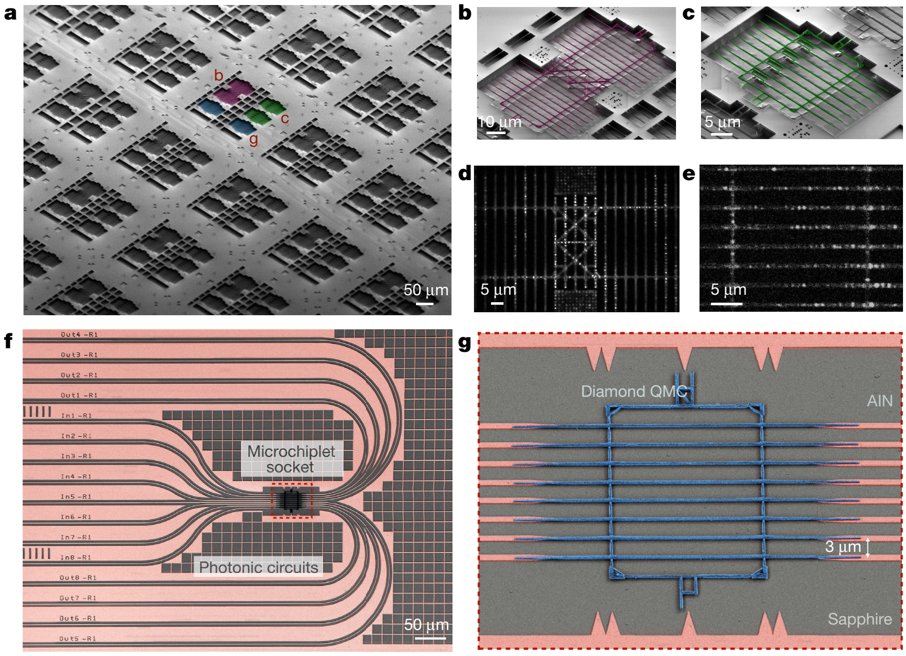


Fig. 2 | Fabrication and integration of QMC with integrated photonics. **a**, SEM overview of the parent diamond chip containing over 500 microchiplets for heterogeneous integration. **b**, A 16-channel QMC. **c**, An 8-channel QMC with varying mechanical beam rigidity. **d**, Photoluminescence map of GeV centres (bright spots) in a 16-channel QMC. **e**, Photoluminescence map of SiV centres (bright spots) in a defect-free 8-channel QMC. **f**, An AlN-on-sapphire integrated photonics module that interfaces with the diamond QMC placed in the chiplet socket. **g**, Close-up SEM of the diamond QMC and AlN photonic interfaces.

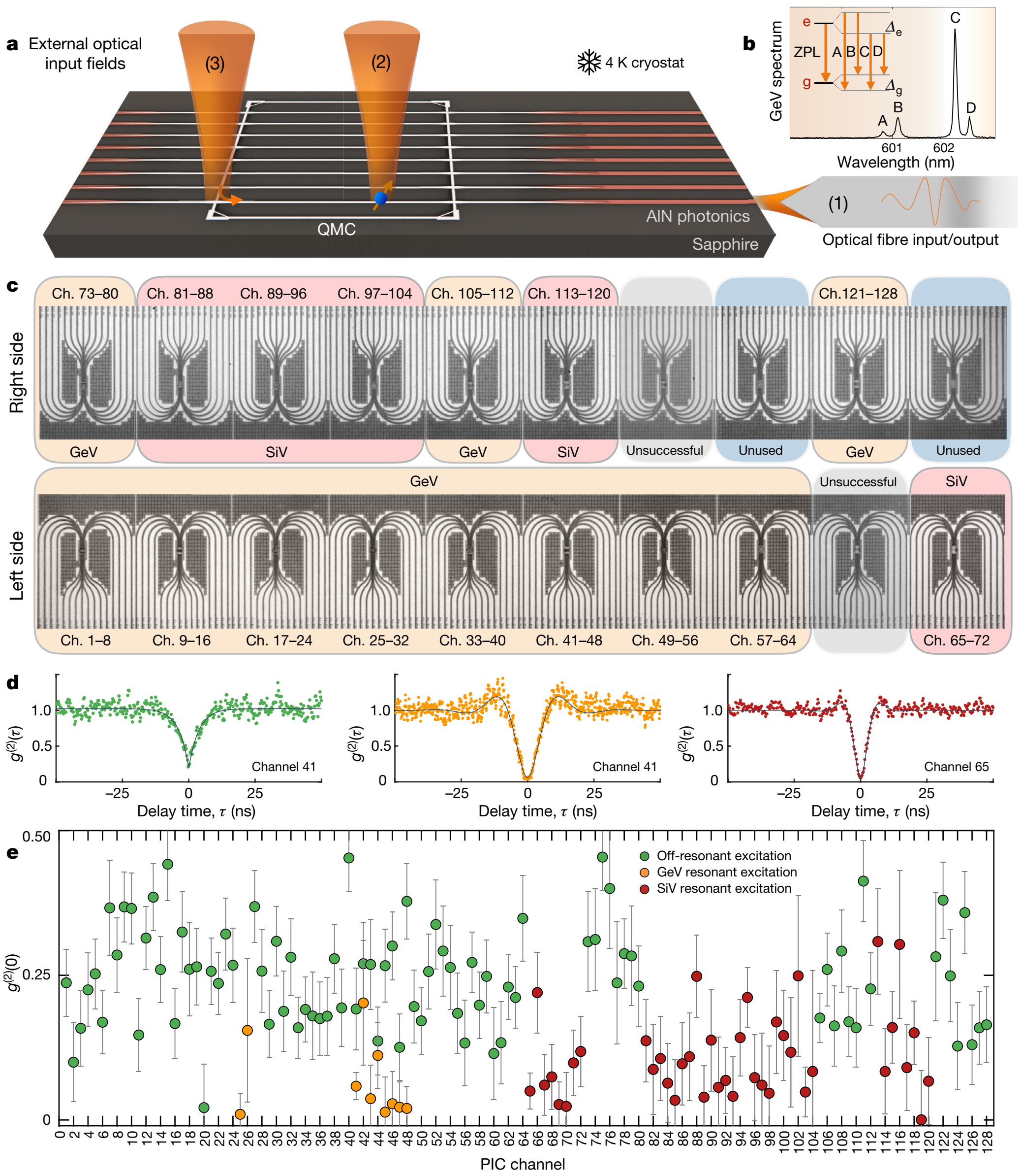


Fig. 3 | Integrated quantum photonics with colour centres. **a**, Experimental setup in a 4-K cryostat showing the input and output optical interfaces (1), (2) and (3). **b**, Energy level and spectrum of a GeV centre, where g, e, Δ_g and Δ_e denote the ground state, excited state, ground-state splitting and excited-state splitting, respectively. Resonant excitation probed transition C, which is the brightest and narrowest line. **c**, Optical image of sixteen QMC-populated microchiplet sockets containing GeV or SiV centres. The 'unsuccessful' modules indicate failed QMC placements. Ch., channel. **d**, Autocorrelation measurements of a single GeV in channel 41 under off-resonant 2-mW, 532-nm excitation (left) and under resonant excitation at 602 nm (middle), and autocorrelation measurement of a single SiV in channel 65 under resonant excitation at 737 nm (right). **e**, Waveguide-coupled single photons from every integrated GeV and SiV channel in the PIC. The error bars indicate fit uncertainties at the 1 s.d. level.

microscope (SEM) image of various suspended chiplets containing 8- or 16-channel waveguide arrays connected by diamond 'trusses', as seen in the close-up SEM images in Fig. 2b, c and Fig. 2g, respectively. Structurally, much larger arrays are fabricable and integrable: we successfully transferred QMCs with as many as 64 waveguide components (Methods). Despite a misalignment between the FIB mask and the QMC patterns, the photoluminescence scans showed that 39% of the 8-channel QMCs are 'defect free' (that is, they have one or more stable colour centre per waveguide) as shown in Fig. 2e (Methods). The defect-free yield of the 16-channel QMCs was lower as these are more susceptible to misalignment, so we did not use them in this study. With improvements in FIB alignment and lithography, as well as targeted fabrication over pre-localized single emitters, an even higher yield should be possible in future work (Methods).

Figure 2f shows one of 20 micro-chiplet sockets connecting 8 input and 8 output waveguide arrays to an 8-channel QMC. We fabricated this PIC on a wafer of single-crystal AlN on a sapphire substrate using EBL and chlorine reactive ion etching[18] (Methods). AlN on sapphire is a suitable platform for linear and nonlinear quantum photonics because of its large bandgap (about 6.2 eV), high material nonlinearities[19,28–30] and low narrowband background fluorescence in the spectrum (600–760 nm) of GeV and SiV centres[18]. Using piezo-controlled micromanipulators, we transferred QMCs into the microchiplet sockets with a placement success rate of 90%. The diamond waveguides (width 340 nm and height 200 nm) transfer light into the AlN waveguides (width 800 nm and height 200 nm) through inverse tapered sections with a simulated efficiency of 97% (98%) at a wavelength of 602 nm (737 nm) (Methods). The SEM image of an assembled device in Fig. 2g shows a transverse

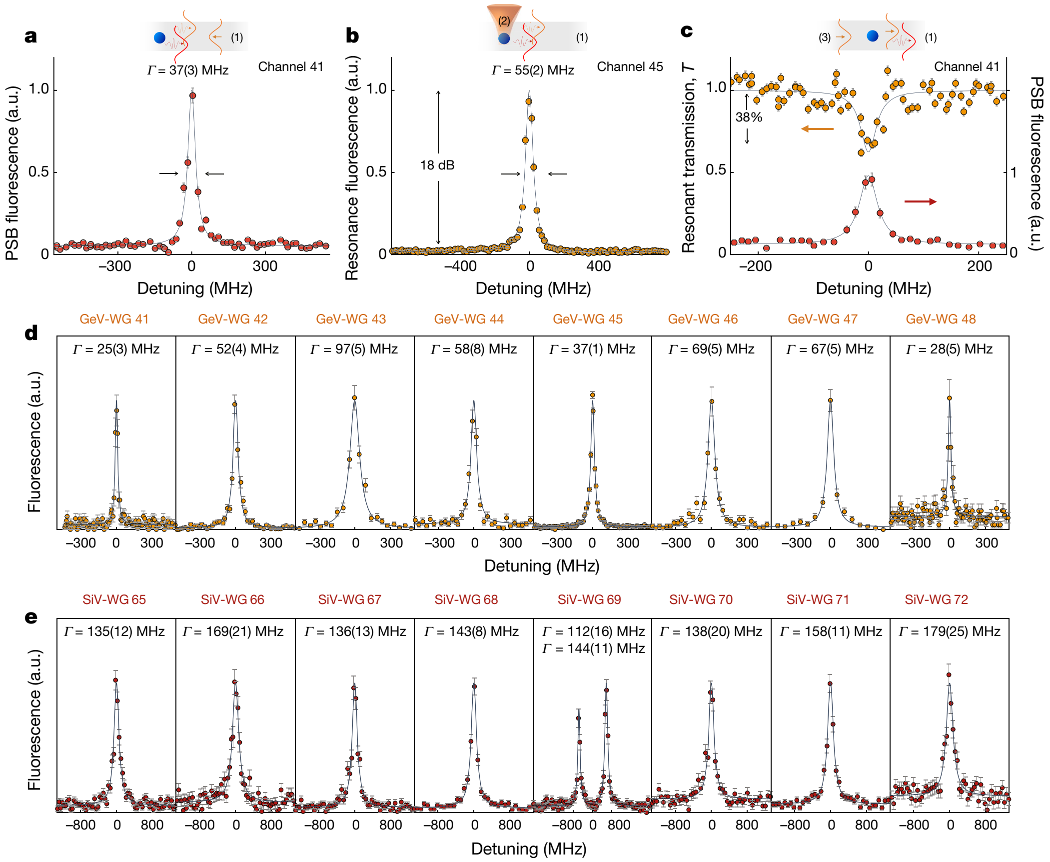


Fig. 4 | Defect-free arrays of optically coherent and efficient waveguide-coupled emitters. a, PLE spectrum (FWHM linewidth $\Gamma = 37(3)$ MHz, indicated by the arrows) of a single GeV in channel 41 with all-fibre excitation and detection of the phonon sideband (PSB) fluorescence routed on-chip via (1). **b**, Excitation via (2) and fluorescence detection via (1). This geometry allows GeV resonance fluorescence detection at least 18 dB above background, without spectral, temporal or polarization filtering. **c**, In transmission, a single GeV centre causes coherent extinction of $\Delta T/T = 38(9)\%$ (orange curve, $\Gamma = 35(15)$ MHz). The red curve shows the PLE spectrum ($\Gamma = 40(5)$ MHz). **d**, PLE spectra of GeV centres in each waveguide (WG) of a characteristic 8-channel GeV QMC, with a mean ± standard deviation linewidth of $\Gamma = 54 \pm 24$ MHz. **e**, PLE spectra of SiVs in an 8-channel SiV QMC, with $\Gamma = 146 \pm 20$ MHz. We interpret the two lines in channel 69 as PLE spectra from two distinct SiV centres ($g^{(2)}(0) = 0.69(7)$ under off-resonant excitation, not shown).

placement error of 38 ± 16 nm. For such typical errors, simulations indicate a drop in coupling efficiency by 10% or 0.46 dB. We find that the transfer of the QMCs is substantially easier than for individual waveguides due to their rigidity and many alignment features. The successful transfer of 16 defect-free chiplets results in a 128-channel photonically integrated quantum emitter chip, as characterized below.

We performed experiments in a closed-cycle cryostat with a base temperature below 4 K, as illustrated in Fig. 3a. The optical fibre labelled (1) couples pump light (fluorescence) to (from) the QMC via the AlN waveguides. A microscope objective also provides optical access to the QMC, for example, to a colour centre (optical interface labelled (2)) or a scattering site (labelled (3)). Figure 3b shows the energy level and emission spectrum of a single GeV when pumped through (2) and collected through (1). Off-resonant excitation using 532-nm light with off-chip pump filtering in this configuration enables the rapid identification of single emitters (indicated by a photon intensity autocorrelation function $g^{(2)}(0) < 0.5$). The left panel in Fig. 3d shows a typical photon antibunching ($g^{(2)}(0) = 0.19(7)$) from a single GeV centre (channel 41) pumped near saturation, without background or detector jitter correction. Under the resonant excitation at 602 nm of transition C (Fig. 3b) of the zero-phonon line (ZPL), the photon purity improves to $g^{(2)}(0) = 0.06(2)$ (middle panel in Fig. 3d). Similarly, in channel 65, we measured antibunched photons with $g^{(2)}(0) = 0.05(3)$ from a single SiV centre under resonant excitation at 737 nm (right panel in Fig. 3d). In all 128 integrated waveguides, shown in Fig. 3c, we identified single GeV and SiV emitters using top excitation (through (2)) and fibre-coupled waveguide collection (through (1)). Their photon statistics are summarized in Fig. 3e.

Next we investigated the optical coherence of a GeV centre using all-fibre spectroscopy. Figure 4a shows the photoluminescence excitation (PLE) spectrum of the channel-41 GeV as we scanned a resonant laser across its ZPL (transition C) with both excitation and detection through the fibre interface (1). Despite the presence of another emitter spectrally detuned by 50 GHz in the same waveguide, resonant excitation allows the selective addressing and readout of single emitters. The measured linewidth of $\Gamma = \Gamma_0 + 2\Gamma_d = 37(3)$ MHz (values in parentheses indicates one standard deviation throughout this work), where Γ_d is the pure dephasing rate of the emitter, is near the lifetime limit $\Gamma_0 = 1/2\pi\tau = 24(2)$ MHz, as obtained from the excited-state lifetime τ (Methods).

The PIC geometry also enables the direct detection of ZPL resonance fluorescence without any spectral, temporal or polarization filtering, even under resonant excitation. Figure 4b shows the resonance fluorescence obtained for top excitation (through (2)) and waveguide collection without filtering in the detection via (1). By polarizing the pump electric field along the waveguide axis to minimize excitation of the transverse electric waveguide mode, this cross-excitation/detection configuration achieves a ZPL intensity 18 dB above the background, comparable to free-space diamond entanglement experiments using cross-polarization and time-gated detection[31].

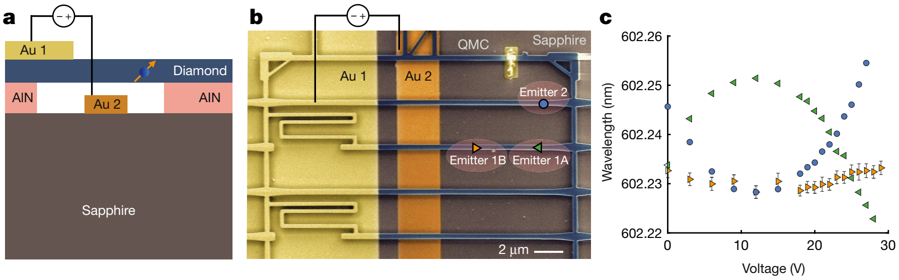


Fig. 5 | Controlling the optical transitions of colour centres on a PIC. a, We applied a d.c. bias between the metal layer Au 1 on diamond and metal Au 2 on the substrate to electrostatically actuate the QMC. **b**, SEM image of the device. In this experiment, we investigated the optical response of emitters 1A, 1B and 2 to strain. **c**, Intra-waveguide emitter 1A and emitter 1B overlap spectrally at 24.5 V. Inter-waveguide overlapping between emitter 2 and emitter 1A (1B) occurs at 2 V (12 V). Error bars for emitter 1A and emitter 2 are smaller than the data points.

According to finite-difference time-domain (FDTD) simulations, an ideal emitter in the optimal configuration has a spontaneous emission coupling efficiency of $\beta = 0.8$ into the diamond waveguide. Experimentally, we measured this efficiency by measuring the transmission of a laser field through a single GeV centre (Fig. 4c). By injecting a laser field through (3) and monitoring the transmission T via (1), we observed an extinction of $1 - T = 0.38(9)$ when on resonance with the GeV centre. This extinction places a lower bound of the emitter-waveguide cooperativity at $C = 0.27(10)$ and $\beta = 0.21(6)$. By accounting for residual line broadening and for the ZPL emission fraction (about 0.6), the dipole-waveguide coupling efficiency is at least 0.55(18); see Methods for other factors that reduce β.

The excellent coherence of the GeV centre in channel 41 is not unique. Figure 4d reports the linewidths of every channel in a characteristic 8-channel GeV diamond chiplet, all measured through the on-chip routing of fluorescence into an optical fibre. We find a mean ± standard deviation normalized linewidth of $\Gamma/\Gamma_0 = 1.7 \pm 0.7$, with GeV channels 41, 45 and 48 exhibiting lifetime-limited values of 1.0(2), 0.9(1) and 1.0(2), respectively. From these measurements, we also obtained the inhomogeneous ZPL transition frequency distribution of 85 GHz. In these PLE measurements, we averaged each spectrum over about 5 min (5,000 experiments), demonstrating the emitters' long-term stability after heterogeneous integration. Similarly, as shown in Fig. 4e, we also find uniformly narrow lines from SiV centres across a QMC, with linewidths within a factor of $\Gamma/\Gamma_0 = 1.6 \pm 0.2$ from SiV centres in bulk diamond[32], and with an inhomogeneous frequency distribution of 30 GHz.

To overcome the inhomogeneous spread in transition frequencies, we implemented a strain-tuning scheme using the electrical layers in our PIC. The fabricated device (Fig. 5a, b) uses a QMC that consists of waveguides with different lengths and beam rigidities (Extended Data Fig. 5). Strain is applied by a capacitive actuator consisting of one gold electrode (Au 1) on the QMC layer, separated transversely by 1.5 µm from a gold ground plane (Au 2) on the sapphire substrate. A bias voltage deforms the waveguide so the associated strain modifies the orbital structures and the optical transitions of embedded colour centres[33–35]. This device geometry enables tuning ranges up to 100 GHz, which is larger than the inhomogeneous distribution and only limited by stiction between the QMC and the substrate (Methods). Owing to differences in dipole positions and orientations, we can spectrally overlap the optical transitions of, for example, emitters 1A and 1B in one waveguide at a bias of 24.5 V, as shown in Fig. 5c. Alternatively, they can also be selectively aligned with that of emitter 2, initially detuned by about 10 GHz in another waveguide channel, at distinct voltages. During strain tuning, we did not observe degradation in the linewidths in PLE scans lasting 3 min (Extended Data Fig. 8); this long-term stability remained within 150 MHz over 3 h of continuous measurement without feedback and unchanged up to a tuning of about 6.8 GHz (Extended Data Fig. 9).

Although not demonstrated here, an array of electrodes could provide closed-loop tuning[35] on each waveguide-coupled emitter to generate indistinguishable photons for Hong–Ou–Mandel interference[36] using on-chip beamsplitters. On the basis of the emitter linewidths measured here over minutes without feedback, we estimate high-visibility interference of 0.9 (0.8) for stable GeV (SiV) emitters such as those in channel 41 (69) and a visibility of 0.58 ± 0.24 (0.63 ± 0.07) when averaging over all emitters in Fig. 4d, e. The optical coherence and photon indistinguishability, which is critical for entangling operations, can be improved, for example, through the Purcell effect by coupling to photonic cavities[37].

The large-scale integration of artificial atoms with photonics extends to a wide range of nanophotonic devices, in particular, high-quality-factor diamond photonic crystal cavities[26,27,38] and other optically active spins[2] such as NV centres[17,31], emerging diamond group-IV quantum memories[39], quantum dots[40] and rare-earth ion dopants[41,42]. The advances reported in this work should therefore encourage further integration of photonic and electronic components for large-scale quantum-information-processing applications such as multiplexed quantum repeaters or modular quantum computers based on solid-state spins[7–12,31,43]. Key components have already been individually demonstrated, including photonic switch arrays and beamsplitter meshes[44–46] for reconfigurable qubit connectivity and heralded spin entanglement, AlN-based high-speed electro-optic modulators[19] and SNSPDs[20], and custom complementary metal–oxide–semiconductor electronics[47,48] for colour centre spin control and low-latency processing. As PIC applications ranging from optical communications[49] to phased-array light detection and ranging[50] to machine learning accelerators[44] are pushing systems beyond thousands of optical components, the high-yield integration with arrays of high-quality artificial atoms provides a basis to extend these scaling gains to quantum information processing with spins and photons.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2441-3.

1. Wehner, S., Elkouss, D. & Hanson, R. Quantum internet: a vision for the road ahead. *Science* **362**, eaam9288 (2018).
2. Awschalom, D. D., Hanson, R., Wrachtrup, J. & Zhou, B. B. Quantum technologies with optically interfaced solid-state spins. *Nat. Photon.* **12**, 516–527 (2018).
3. Atatüre, M., Englund, D., Vamivakas, N., Lee, S.-Y. & Wrachtrup, J. Material platforms for spin-based photonic quantum technologies. *Nat. Rev. Mater.* **3**, 38–51 (2018).
4. Humphreys, P. C. et al. Deterministic delivery of remote entanglement on a quantum network. *Nature* **558**, 268–273 (2018); correction **562**, E2 (2018).
5. Bradley, C. E. et al. A ten-qubit solid-state spin register with quantum memory up to one minute. *Phys. Rev. X* **9**, 031045 (2019).
6. Bhaskar, M. K. et al. Experimental demonstration of memory-enhanced quantum communication. *Nature* **580**, 60–64 (2020).

7. Muralidharan, S. et al. Optimal architectures for long distance quantum communication. *Sci. Rep.* **6**, 20463 (2016).
8. Lo Piparo, N., Munro, W. J. & Nemoto, K. Quantum multiplexing. *Phys. Rev. A* **99**, 022337 (2019).
9. Nemoto, K. et al. Photonic architecture for scalable quantum information processing in diamond. *Phys. Rev. X* **4**, 031022 (2014).
10. Monroe, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
11. Nickerson, N. H., Fitzsimons, J. F. & Benjamin, S. C. Freely scalable quantum technologies using cells of 5-to-50 qubits with very lossy and noisy photonic links. *Phys. Rev. X* **4**, 041041 (2014).
12. Choi, H., Pant, M., Guha, S. & Englund, D. Percolation-based architecture for cluster state creation using photon-mediated entanglement between atomic memories. *npj Quantum Inf*. **5**, 104 (2019).
13. Kim, J.-H. et al. Hybrid integration of solid-state quantum emitters on a silicon photonic chip. *Nano Lett*. **17**, 7394–7400 (2017).
14. Elshaari, A. W. et al. On-chip single photon filtering and multiplexing in hybrid quantum photonic circuits. *Nat. Commun*. **8**, 379 (2017).
15. Osada, A. et al. Strongly coupled single-quantum-dot–cavity system integrated on a CMOS-processed silicon photonic chip. *Phys. Rev. Appl*. **11**, 024071 (2019).
16. Najafi, F. et al. On-chip detection of non-classical light by scalable integration of single-photon detectors. *Nat. Commun*. **6**, 5873 (2015).
17. Mouradian, S. L. et al. Scalable integration of long-lived quantum memories into a photonic circuit. *Phys. Rev. X* **5**, 031009 (2015).
18. Lu, T.-J. et al. Aluminum nitride integrated photonics platform for the ultraviolet to visible spectrum. *Opt. Express* **26**, 11147–11160 (2018).
19. Xiong, C., Pernice, W. H. P. & Tang, H. X. Low-loss, silicon integrated, aluminum nitride photonic circuits and their use for electro-optic signal processing. *Nano Lett*. **12**, 3562–3568 (2012).
20. Zhu, D. et al. Superconducting nanowire single-photon detector on aluminum nitride. In *Conference on Lasers and Electro-Optics* FTu4C.1 (Optical Society of America, 2016).
21. Sukachev, D. D. et al. Silicon-vacancy spin qubit in diamond: a quantum memory exceeding 10 ms with single-shot state readout. *Phys. Rev. Lett*. **119**, 223602 (2017).
22. Bhaskar, M. K. et al. Quantum nonlinear optics with a germanium-vacancy color center in a nanoscale diamond waveguide. *Phys. Rev. Lett*. **118**, 223603 (2017).
23. Becker, J. N. & Becher, C. Coherence properties and quantum control of silicon vacancy color centers in diamond. *Phys. Status Solidi A* **214**, 1770170 (2017).
24. Siyushev, P. et al. Optical and microwave control of germanium-vacancy center spins in diamond. *Phys. Rev. B* **96**, 081201(R) (2017).
25. Nguyen, C. T. et al. Quantum network nodes based on diamond qubits with an efficient nanophotonic interface. *Phys. Rev. Lett*. **123**, 183602 (2019).
26. Mouradian, S., Wan, N. H., Schröder, T. & Englund, D. Rectangular photonic crystal nanobeam cavities in bulk diamond. *Appl. Phys. Lett*. **111**, 021103 (2017).
27. Wan, N. H., Mouradian, S. & Englund, D. Two-dimensional photonic crystal slab nanocavities on bulk single-crystal diamond. *Appl. Phys. Lett*. **112**, 141102 (2018).
28. Lueng, C. M., Chan, H. L. W., Surya, C. & Choy, C. L. Piezoelectric coefficient of aluminum nitride and gallium nitride. *J. Appl. Phys*. **88**, 5360–5363 (2000).
29. Guo, X., Zou, C.-L., Jung, H. & Tang, H. X. On-chip strong coupling and efficient frequency conversion between telecom and visible optical modes. *Phys. Rev. Lett*. **117**, 123902 (2016).
30. Jung, H., Xiong, C., Fong, K. Y., Zhang, X. & Tang, H. X. Optical frequency comb generation from aluminum nitride microring resonator. *Opt. Lett*. **38**, 2810–2813 (2013).
31. Bernien, H. et al. Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497**, 86–90 (2013).
32. Rogers, L. J. et al. Multiple intrinsically identical single-photon emitters in the solid state. *Nat. Commun*. **5**, 4739 (2014).
33. Meesala, S. et al. Strain engineering of the silicon-vacancy center in diamond. *Phys. Rev. B* **97**, 205444 (2018).
34. Maity, S. et al. Spectral alignment of single-photon emitters in diamond using strain gradient. *Phys. Rev. Appl*. **10**, 024050 (2018).
35. Machielse, B. et al. Quantum interference of electromechanically stabilized emitters in nanophotonic devices. *Phys. Rev. X* **9**, 031022 (2019).
36. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett*. **59**, 2044–2046 (1987).
37. Grange, T. et al. Reducing phonon-induced decoherence in solid-state single-photon sources with cavity quantum electrodynamics. *Phys. Rev. Lett*. **118**, 253602 (2017).
38. Mouradian, S. L. & Englund, D. A tunable waveguide-coupled cavity design for scalable interfaces to solid-state quantum emitters. *APL Photon*. **2**, 046103 (2017).
39. Bradac, C., Gao, W., Forneris, J., Trusheim, M. E. & Aharonovich, I. Quantum nanophotonics with group IV defects in diamond. *Nat. Commun*. **10**, 5625 (2019); correction **11**, 360 (2020).
40. Lodahl, P., Mahmoodian, S. & Stobbe, S. Interfacing single photons and single quantum dots with photonic nanostructures. *Rev. Mod. Phys*. **87**, 347–400 (2015).
41. Zhong, T. et al. Optically addressing single rare-earth ions in a nanophotonic cavity. *Phys. Rev. Lett*. **121**, 183603 (2018).
42. Dibos, A. M., Raha, M., Phenicie, C. M. & Thompson, J. D. Atomic source of single photons in the telecom band. *Phys. Rev. Lett*. **120**, 243601 (2018).
43. Bersin, E. et al. Individual control and readout of qubits in a sub-diffraction volume. *npj Quantum Inf*. **5**, 38 (2019).
44. Harris, N. C. et al. Linear programmable nanophotonic processors. *Optica* **5**, 1623–1631 (2018).
45. Taballione, C. et al. 8×8 reconfigurable quantum photonic processor based on silicon nitride waveguides. *Opt. Express* **27**, 26842–26857 (2019).
46. Seok, T. J., Quack, N., Han, S., Muller, R. S. & Wu, M. C. Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers. *Optica* **3**, 64–70 (2016).
47. Kim, D. et al. A CMOS-integrated quantum sensor based on nitrogen–vacancy centres. *Nat. Electron*. **2**, 284–289 (2019).
48. Patra, B. et al. Cryo-CMOS circuits and systems for quantum computing applications. *IEEE J. Solid-State Circuits* **53**, 309–321 (2018).
49. Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018); **560**, E4 (2018).
50. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).

Methods

Ion implantation

Extended Data Fig. 1 summarizes the fabrication and integration processes. First, we relieved the strained surface of the single-crystal diamond plate (Element6) by plasma etching the first 10 μm of diamond in Ar/Cl_2, followed by another 5 μm etching in pure oxygen plasma. We used an FIB[51] tool at the Ion Beam Laboratory (Sandia National Laboratories) to implant Ge ions (spot size of about 35 nm × 43 nm) and Si ions (spot size of about 50 nm × 45 nm) at an effective areal dose of 2×10^{11}–6×10^{11} ions per cm^2 and 4.5×10^{11}–9×10^{11} ions per cm^2, respectively. The Ge (Si) ion energy is 200 keV (170 keV), which corresponds to an implantation depth of 74 ± 12 nm (113 ± 19 nm) from stopping and range of ions in matter (SRIM) simulations[52]. After implantation, we annealed the devices at 1,200 °C in an ultrahigh vacuum furnace. Finally, we cleaned the diamond in a boiling mixture of 1:1:1 sulfuric acid, nitric acid and perchloric acid.

Conversion yield of GeV and SiV centres

We analysed the conversion yields of GeV and SiV centres by counting the absence of fluorescent spots in our implantation region (1-μm pitch, square grid) using photoluminescence microscopy. A Poisson distribution $P(k)$, with mean number of colour centres λ and number of observed emitters per spot k, models the stochastic emitter creation process. From the mean $\lambda = -\log(P(0))$ and our implantation dose, we estimate the conversion yield of GeV (SiV) centres to be about 1.9% (3.2%).

Registration of emitters using optical localization

We located and mapped the fabricated quantum emitters relative to prefabricated alignment markers using a wide-field and confocal scanning microscope as shown previously[53]. To demonstrate the microchiplet principle in this study, we registered the qubit grid, rather than each emitter's location. In particular, we determined the global displacement of the emitter grid from the implantation process and used this offset in our subsequent electron-beam lithography of QMCs. We anticipate that the targeted placement of devices over pre-localized emitters[53–57] will increase the yield of emitter-coupled QMCs. Such approaches may also be critical for reducing the proximity of nanofabricated surfaces to emitters and for coupling to photonic crystal nanocavities.

Quantum microchiplet fabrication

After ion implantation and optical registration, we used a quasi-isotropic diamond etching recipe[26,27] to fabricate suspended QMCs. In particular, we deposited 180 nm of silicon nitride (SiN) hard mask using plasma-enhanced chemical vapour deposition. We patterned the SiN hard mask using a ZEP-520A electron-beam resist with ESpacer conductive polymer and CF_4 reactive-ion etching. Subsequently, we used inductively coupled reactive-ion etching to transfer the pattern from SiN into the diamond layer. Following oxygen etching of the diamond, we deposited 15 nm of conformal alumina via atomic layer deposition. After a brief breakthrough etch of alumina, we etched the chip in zero-bias oxygen plasma to isotropically undercut the diamond QMCs. Finally, we removed the SiN and alumina masks in hydrofluoric acid. We again annealed the device at 1,200 °C using the above ultrahigh-vacuum, high-temperature annealing recipe, followed by a clean in a boiling mixture of 1:1:1 sulfuric acid, nitric acid and perchloric acid.

AlN photonics

AlN is a large-bandgap material (about 6.2 eV) that is suitable for linear optics, nonlinear optics and optomechanics, with an electro-optic coefficient of $r_{33} = 1\,\mathrm{pm\,V^{-1}}$ (ref. [19]), second-order optical nonlinear susceptibility of $\chi^{(2)} = 4.7\,\mathrm{pm\,V^{-1}}$ (ref. [29]), a third-order optical nonlinearity (Kerr) coefficient of $n_2 \approx 2.5 \times 10^{-15}\,\mathrm{cm^2\,W^{-1}}$ (ref. [30]) and a piezoelectric coefficient of $d_{33} \approx 5\,\mathrm{pm\,V^{-1}}$ (ref. [28]). Here we used a wafer of 200-nm-thick single-crystal AlN on a sapphire substrate (MSE Supplies, grown by hydride vapour phase epitaxy). Before processing of the AlN PIC, we patterned gold alignment markers to use for alignment between the photonic layer and the metal layers for strain tuning. We defined the AlN photonic circuitry using EBL (ZEP-520A electron-beam resist and ESpacer conductive polymer) and chlorine-based inductively coupled plasma reactive-ion etching)[18]. Then, S1813 photoresist served as a protective layer for mechanical edge polishing. We then diced the chip using an automatic dicing saw (DISCO DAD-3240). We polished the chip to produce optical-grade facets for edge coupling (Allied Multi-Prep Polishing System 8). Finally, sonication in *N*-methyl-2-pyrrolidone removed the S1813 protective layer and debris caused by dicing and mechanical polishing.

Metal layers

The fabrication of the metal electrodes and contact pads on top of the PIC substrate immediately followed the patterning of the thin-film AlN and preceded the chip dicing and edge polishing. The PIC substrate metal layer was defined by lift-off of 50-nm Au on top of 5-nm Ti using a single layer of A6 950K poly(methyl methacrylate) electron-beam resist (450 nm thick), which was aligned relative to the AlN PIC with metal alignment markers. Then, the fabrication of the AlN photonic circuitry proceeded to dicing and polishing, followed by integration of the QMC. After pick-and-place transfer of the QMC to the microchiplet socket, we used a targeted electron-beam metal deposition process to place platinum on the periphery of the QMC for electrical connection (FEI Helios NanoLab 600 DualBeam). This process also locked the QMC into place before resist spin-coating. Finally, we defined the metal electrode layer on top of the QMC by lift-off of 15-nm Au on 5-nm Ti using a single layer of A11 950K poly(methyl methacrylate) (2 μm thick).

Yield of defect-free microchiplets

Using photoluminescence spectroscopy, we investigated the occurrence of defect-free 8-channel QMCs, as summarized in Extended Data Fig. 2. From this histogram, we estimate the probability of creating defect-free QMCs to be 39%. We note that this success probability depends on a variety of factors, including the alignment accuracy of the FIB implantation, the relative calibration between EBL and FIB, as well as the optical registration process. By deterministically placing each element of the QMC over pre-localized emitters, it should be possible to boost the yield towards unity, allowing hundreds or thousands of quantum channels per chiplet.

Pick-and-place transfer process

We used piezo-controlled three-axis and rotation stages to align the QMC with the PIC[17]. In addition to the AlN waveguides, the QMC also rests on top of multiple small AlN pedestals to prevent bowing of the diamond structures and stiction with the underlying sapphire substrate. In the case of an inaccurate placement, both the QMC and socket can be reused simply by picking the QMC and re-attempting the placement process. Experimentally, we have transferred a variety of arrays, ranging from single-channel devices all the way to 64-channel QMCs. We expect computer-controlled placement and self-alignment locking features to improve the transfer rate and to potentially fully automate the process.

Experimental setup

We used a closed-cycle helium cryostat with a base temperature of 4 K (Montana Instruments) with a top-access microscope objective (Mitutoyo 100× ULWD, numerical aperture (NA) of 0.55). We used three-axis nanoposition steppers (Attocube ANP-x,z-50) and scanners (Attocube ANS-x,z-50) for edge coupling of optical fibres (lensed fibre with a spot size of 0.8 μm at 633 nm or a Nufern UHNA3 fibre) to the PIC. For photoluminescence (PLE) spectroscopy, we filtered the fibre-coupled fluorescence in free space using bandpass filters—Semrock FF01-605/15 (FF01-647/57) for GeV centres and FF01-740/13 (FF01-775/46) for SiV centres. We off-resonantly pumped GeV (SiV) using 532-nm (660-nm) lasers. Resonant excitation was achieved using a tunable laser (MSquared SolsTiS with an external mixing module). For PLE, we used acousto-optic modulators to excite GeV centres with a resonant pulse and an optional 532-nm charge repump pulse. For SiV centre experiments, we did not gate the resonant and repump optical pulses. In the resonance fluorescence detection experiment (Fig. 4b), we placed a half-wave plate before channel 2 to minimize laser coupling into the waveguide mode. To measure the excited-state lifetime of single emitters, we used time-correlated single-photon counting (PicoHarp 300) and a pulsed laser source (SuperK, filtered to 532 ± 20 nm). We fitted the lifetime curves of the emitters in Fig. 4d with biexponential terms to account for fast decay and the slower fluorescence decay time constant. For strain-tuning experiments, we used a programmable voltage source (Keithley 2400) and observed negligible leakage currents (less than 0.2 nA) for all applied voltages in this experiment (up to 35 V).

Spontaneous emission β-factor and dipole coupling with the waveguide mode

The extinction in the resonant transmission spectrum arises from the interference between the scattered and incoming optical fields, and its depth depends on the dipole-waveguide coupling $\beta = \Gamma_{\mathrm{wg}}/(\Gamma_{\mathrm{wg}} + \Gamma')$, where Γ_{wg} is the emission rate into the waveguide mode and Γ' the decay rate into all other channels. For the measurement in Fig. 4c, we first characterized the saturation response of the emitter when excited via (3). At the low-excitation limit, the cooperativity C can be extracted from $T \approx (1-\beta)^2 = (1+C)^{-2}$. By also accounting for line broadening of $2\Gamma_{\mathrm{d}}/\Gamma_0 = 0.33(14)$, we determined β via[58] $T = 1 - \frac{(\beta - 2)\beta}{(1 + 2\Gamma_{\mathrm{d}}/\Gamma)(1+S)}$, which reduces to the usual expression[59] $T \approx (1-\beta)^2$ in the absence of broadening and far from saturation $S \ll 1$. In this experiment, we operated at $S \approx 10^{-2}$ and all errors denote the fit or propagated uncertainties. The discrepancy of the experimental $\beta = 0.21(6)$ (0.55(18) after correcting for broadening and a ZPL branching ratio of 0.6) with the simulated $\beta = 0.8$ using the three-dimensional (3D) FDTD method (Lumerical) arises from three possible sources: (1) angular and positional misalignment of the dipole in the waveguide; (2) a finite population in the upper ground state and emission into transition D; and (3) possible non-radiative processes.

Diamond–PIC coupling

Extended Data Fig. 3a, b shows the normalized electric |**E**| field of 602-nm (737-nm)-wavelength transverse electric light coupling from the diamond waveguide (340 nm × 200 nm) to the bottom AlN waveguide (800 nm × 200 nm), calculated using the 3D FDTD method. The light transfers adiabatically via tapered sections in the diamond waveguide and AlN waveguide. Here the coupling region is 9 μm long, with a diamond taper length of 8 μm and AlN taper length of 5 μm. The top insets show 2D transverse cross-sections of the light propagation. The cross-sections at $y = -10$ μm and $y = 1$ μm correspond to the fundamental transverse electric mode of the diamond waveguide and AlN-on-sapphire waveguide, respectively. The cross section at $y = -5$ μm ($y = -6$ μm) is the point where half of the light launched from the diamond waveguide is transferred to the AlN waveguide at 602-nm (737-nm) wavelength. The light from the diamond waveguide couples to the AlN waveguide with 97% (98%) efficiency at these wavelengths, with all of the light coupling preferentially to the AlN fundamental transverse electric mode and negligible coupling to higher-order modes. This optimized device geometry was determined by optimizing for the coupling efficiency from the fundamental transverse electric mode of the diamond waveguide to the fundamental transverse electric mode of the AlN while sweeping the diamond taper length, the AlN taper length, and the overlap region between the diamond and AlN waveguides. In Fig. 2g, we showed a typical transverse placement error of 38 ± 16 nm for our transfer placement of the QMC to the microchiplet socket; in simulation, this displacement corresponds to a decrease of the coupling efficiency to 93% (89%) at 602-nm (737-nm) wavelength. Hence, we have a 0.46-dB tolerance in the coupling efficiency within our transfer placement accuracy. By directly measuring the PIC–diamond–PIC transmission efficiency, we found the interlayer coupling efficiency to be greater than 34%, which was lower than simulations probably due to scattering at the interfaces and the QMC cross-junctions.

PIC–fibre coupling

We couple laser and photoluminescence to and from AlN-on-sapphire waveguides using lensed fibres (Nanonics Imaging, SM-630 with spot size 0.8 ± 0.3 μm and working distance 4 ± 1 μm) for cryostat experiments and ultrahigh NA fibres (UNHA3) for room-temperature experiments. Under our single-mode operation at 602–737 nm, the in-coupling efficiency is the same as the out-coupling efficiency of AlN waveguide to lensed fibre, which we find to be 51–57% using the 3D FDTD method. In practice, the PIC–fibre coupling efficiency, which we find to be about 11% in our devices, is sensitive to the edge coupler polishing quality. For the high-NA fibre, which is multimode at our wavelengths of interest, we find the numerical out-coupling efficiency to the fundamental fibre mode to be 25% (34%) at 602 nm (737 nm); there is also 1% (3%) coupling into higher-order modes.

System efficiency η

Extended Data Fig. 4a shows the response from an idealized emitter system, fitted to $F = F_{\mathrm{sat}}P/(P_{\mathrm{sat}} + P) + cP$, where P is the continuous-wave 532-nm excitation pump power, cP is the linear background, $P_{\mathrm{sat}} = 1.2$ mW is the saturation power, F is the measured ZPL fluorescence at the detector and $F_{\mathrm{sat}} = 1.11$ megacounts per second (Mcps). Extended Data Table 1 shows an average saturated count rate of 0.64 ± 0.36 Mcps from an array of GeV waveguides in a QMC. To independently measure the system efficiency at the detector, η_{system}, we used a pulsed source (SuperK Extreme, 532 ± 20 nm) with a repetition rate of 26 MHz. From the saturation response (Extended Data Fig. 4b), we determined $F_{\mathrm{sat}} = 0.25$ Mcps and $\eta_{\mathrm{system}} = 0.72\%$. This experimentally determined efficiency is within a factor of five from the independently calculated $\eta_{\mathrm{system}} = 0.5\beta\eta_{\mathrm{PIC}}\eta_{\mathrm{fibre}}\eta_{\mathrm{setup}} \approx 2.6\%$, where $(\beta, \eta_{\mathrm{PIC}}, \eta_{\mathrm{fibre}}, \eta_{\mathrm{setup}}) \approx (0.55, 0.34, 0.33, 0.58)$ are the dipole-waveguide, diamond–PIC coupling, PIC–fibre coupling and external setup detection efficiencies, respectively. Here the factor of 0.5 accounts for the present configuration in which we collected the photon emission in one direction only. In these saturation experiments at room temperature, we used a lensed fibre with 2.5-μm spot size at 1,550 nm, which we find to have $\eta_{\mathrm{fibre}} \approx 33\%$. We attribute the discrepancy to the non-unity radiative quantum efficiency of the emitter and deviations in β, η_{PIC} and η_{fibre} from independent measurements based on another device. In the next subsection, we outline methods to improve the system efficiency.

Improving the system efficiency

In our experiments, the uncladded microchip enables the heterogeneous integration of QMC but the mode mismatch between the AlN-on-sapphire waveguide and the lensed (high NA) fibre causes at least 3 dB (5 dB) insertion losses as characterized above. It is possible to improve the mode overlap by cladding the microchip with alumina or with materials with similar refractive indices as the underlying sapphire[18]. In such a scheme, we would taper down the AlN waveguides at the chip facet to better mode-match with the lensed fibre. Our cladded edge coupler design can substantially improve the coupling from the AlN waveguide to the lensed fibre to be 84% (84%) at 602 nm (737 nm), as well as improve the coupling to high NA fibre to be 89% (91%) at 602 nm (737 nm), using the present AlN-on-sapphire material and film thickness. In this design, we matched the mode field diameters and reduced the effective refractive index mismatch between the fundamental transverse electric modes at AlN edge coupler facet and the lensed (high NA) fibre focus spot (facet). Owing to the index mismatch, the light coupling is limited by Fresnel reflections at the waveguide facet, which can be reduced using an index-matching environment. Finally, on-chip reflectors in diamond can increase the photon collection efficiency by a factor of two, and photonic crystal cavities can boost the emission into the waveguide mode.

Strain tuning scheme of QMC on PIC

We introduced different optical responses to our emitter QMC by changing the length of their constituent waveguides. Here we used waveguides of length 20 μm (type I) and length 15 μm (type II). To be compatible with the QMC framework, we included a flexible bridge between type II waveguides and the QMC body (Extended Data Fig. 5a, Fig. 5a). Extended Data Fig. 5b confirms the difference in strain response at 30 V (modelled using COMSOL Multiphysics) between type I and type II waveguides.

Response of optical transitions to strain

We consider single GeV centres (emitter 1A, emitter 1B, emitter 2) indicated in Extended Data Fig. 5a, Fig. 5a, b. Extended Data Fig. 6 plots the spectral response of the optical transition lines up to an applied voltage of 30 V. From the increasing line splitting of the orbital ground states Δ_g, that is, between lines C and D (as well as A and B), we find that emitter 1B is a dipole whose axis lies in the transverse plane[33,34,60] of the waveguide. On the basis of the unidirectional shift of all four lines, emitters 1A and 2 are dipoles oriented in the longitudinal cross-sectional plane of the waveguide[33,34,60]. In particular, the global blueshift of the lines of emitter 1A indicates that it resides in a region with compressive strain (that is, below the neutral axis of the mechanical beam). Conversely, the optical lines of emitter 2 redshifts with applied voltage, indicating that it resides in a region with tensile strain, which is located above the neutral axis of the waveguide. Extended Data Fig. 7 shows the robustness of the strain-tuning mechanism as we repeatedly applied voltages from 10 V to 26 V. Above 30 V, we see over 100 GHz of tuning of the two brightest transitions C and D for emitters 1A and 2; however, we note that in this regime there was hysteresis possibly due to stiction with the underlying gold and substrate about 150 nm and 200 nm away, respectively. Nevertheless, for the purpose here, we were able to spectrally overlap any pair of the three emitters with less than 25 V. Revised electrode, QMC and/or PIC designs in future microchips should be able to extend the spectral shift of individually tunable waveguides. We note that the small ‘pull in’ voltage in our experiment appears earlier than it does in simulation (over 250 V)—possibly due to the surface conductivity of diamond[60,61].

Stability of optical transitions

We investigated the optical stability of an emitter during spectral tuning via strain. Here we monitored a GeV in another chiplet with an identical electrode configuration (the earlier device used for Fig. 5 and Extended Data Figs. 5–7 was no longer available due to an accident). Extended Data Fig. 8 shows the centre frequency shift (bottom) of the ZPL transition and its linewidth (top) as a function of voltage. In these PLE linescans under strain, we averaged each spectrum over 2,000 experiments (about 3 min) and did not observe substantial degradation in the linewidth. We then tracked the ZPL at various voltage biases under repeated PLE measurements over 3 h. Extended Data Fig. 9 shows the long-term ZPL stability to within 150 MHz for spectral tuning up to 6.8 GHz. At higher tuning ranges, the linewidths were unchanged but there was an increase in spectral diffusion of the centre frequency, probably due to an induced permanent dipole moment that increased susceptibility to charge fluctuations[60,62]. Nevertheless, at a tuning of 20 GHz, the FWHM of the inhomogeneous distribution remained under 250 MHz, which is within a factor of four of the initial linewidths of about 60 MHz.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. The data that support the findings of this study are also openly available in figshare at https://doi.org/10.6084/m9.figshare.11874291.

51. Schröder, T. et al. Scalable focused ion beam creation of nearly lifetime-limited single quantum emitters in diamond nanostructures. *Nat. Commun.* **8**, 15376 (2017).
52. Ziegler, J. F., Ziegler, M. D. & Biersack, J. P. SRIM – the stopping and range of ions in matter (2010). *Nucl. Instrum. Methods Phys. Res. B* **268**, 1818–1823 (2010).
53. Wan, N. H. et al. Efficient extraction of light from a nitrogen-vacancy center in a diamond parabolic reflector. *Nano Lett.* **18**, 2787–2793 (2018).
54. Badolato, A. et al. Deterministic coupling of single quantum dots to single nanocavity modes. *Science* **308**, 1158–1161 (2005).
55. Gazzano, O. et al. Bright solid-state sources of indistinguishable single photons. *Nat. Commun.* **4**, 1425 (2013).
56. Gschrey, M. et al. In situ electron-beam lithography of deterministic single-quantum-dot mesa-structures using low-temperature cathodoluminescence spectroscopy. *Appl. Phys. Lett.* **102**, 251113 (2013).
57. Sapienza, L., Davanço, M., Badolato, A. & Srinivasan, K. Nanoscale optical positioning of single quantum dots for bright and pure single-photon emission. *Nat. Commun.* **6**, 7833 (2015).
58. Thyrrestrup, H. et al. Quantum optics with near-lifetime-limited quantum-dot transitions in a nanophotonic waveguide. *Nano Lett.* **18**, 1801–1806 (2018).
59. Chang, D. E., Sørensen, A. S., Demler, E. A. & Lukin, M. D. A single-photon transistor using nanoscale surface plasmons. *Nat. Phys.* **3**, 807–812 (2007).
60. Sohn, Y.-I. et al. Controlling the coherence of a diamond spin qubit through its strain environment. *Nat. Commun.* **9**, 2012 (2018).
61. Maier, F., Riedel, M., Mantel, B., Ristein, J. & Ley, L. Origin of surface conductivity in diamond. *Phys. Rev. Lett.* **85**, 3472–3475 (2000).
62. Sipahigil, A. et al. Indistinguishable photons from separated silicon-vacancy centers in diamond. *Phys. Rev. Lett.* **113**, 113602 (2014).

Acknowledgements The focused ion beam implantation work was performed at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the US Department of Energy (DOE) Office of Science. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the US DOE’s National Nuclear Security Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the US DOE or the United States Government. This work made use of the Shared Experimental Facilities supported in part by the MRSEC Program of the National Science Foundation (NSF) under award number DMR - 1419807. We thank D. Perry for providing the focused ion beam implantation at Sandia National Laboratories, and D. Zhu and C. Peng for assistance with wire bonding. N.H.W. acknowledges support from the Army Research Laboratory (ARL) Center for Distributed Quantum Information (CDQI) programme W911NF-15-2-0067. T.-J. L. acknowledges support from the Department of Defense (DOD) National Defense Science and Engineering Graduate Fellowship (NDSEG) as well as the Air Force Research Laboratory RITA programme FA8750-16-2-0141. K.C.C. acknowledges funding support by the NSF Graduate Research Fellowships Program and ARL CDQI. M.P.W. acknowledges support from the NSF Center for Integrated Quantum Materials (CIQM), NSF grant number DMR-1231319. M.T. acknowledges support by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the Massachusetts Institute of Technology, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US DOE and the Office of the Director of National Intelligence. L.D.S acknowledges support from the Under Secretary of Defense for Research and Engineering administered through the MIT Lincoln Laboratory Technology Office. E.A.B. was supported by a NASA Space Technology Research Fellowship and the NSF Center for Ultracold Atoms (PHY-1734011). I.B.H is supported by the DOE ‘Photonics at Thermodynamic Limits’ Energy Frontier Research Center under grant DE-SC0019140. S.L.M was supported by the NSF EFRI ACQUIRE programme EFMA-1641064. I.R.C. acknowledges funding support from the DOD NDSEG Fellowship, NSF award DMR-1747426, and the NSF EFRI ACQUIRE programme EFMA-1641064. D.E. acknowledges partial support from the MITRE Quantum Moonshot initiative.

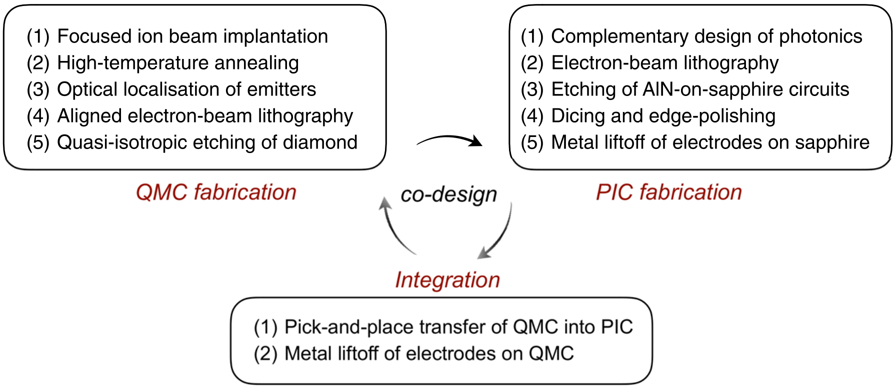
Author contributions N.H.W., T.-J.L. and D.E. conceived the experiments and wrote the manuscript. N.H.W. and T.-J.L. fabricated the devices, performed the experiments, and analysed the data, with fabrication assistance from K.C.C. and experimental assistance from M.P.W., M.E.T., L.D.S., E.A.B., I.B.H., S.L.M., and I.R.C. E.S.B. performed ion implantation. D.E. supervised the project. All authors discussed the results and contributed to the manuscript.

Competing interests The authors declare no competing interests.

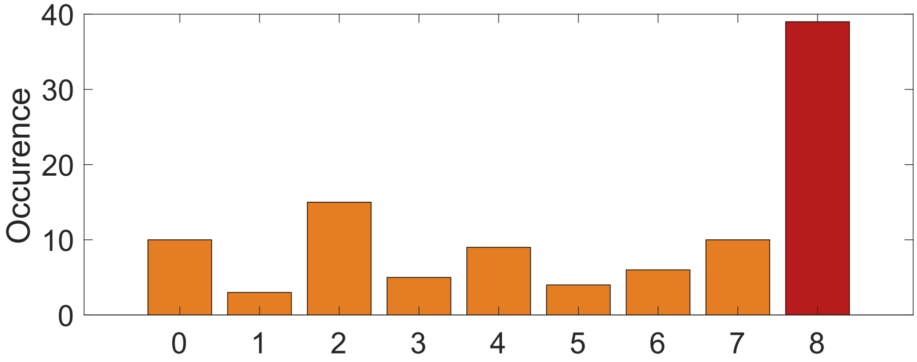
Additional information

Correspondence and requests for materials should be addressed to N.H.W., T.-J.L. or D.E.

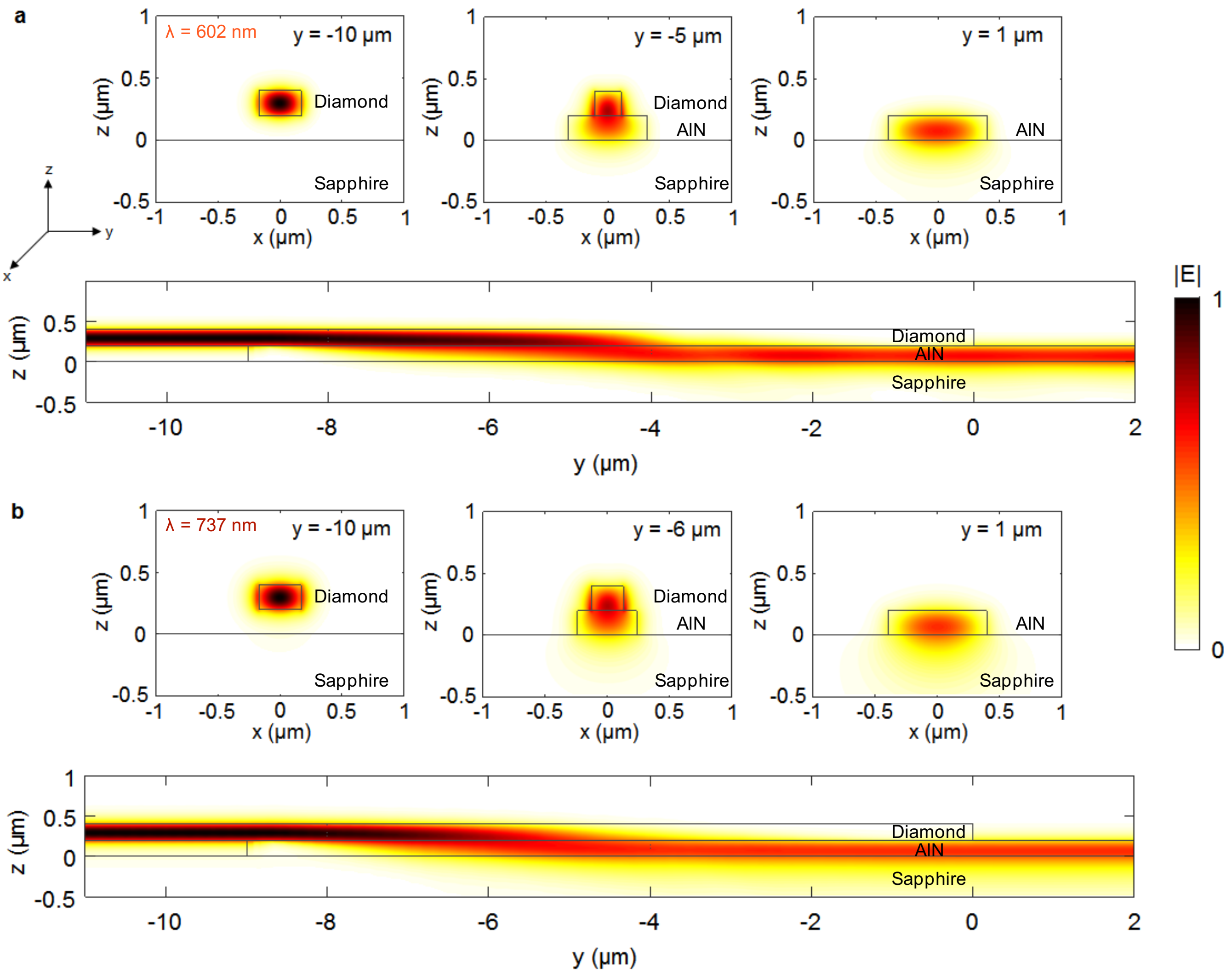
Peer review information *Nature* thanks Wolfram Pernice, Jennifer Choy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



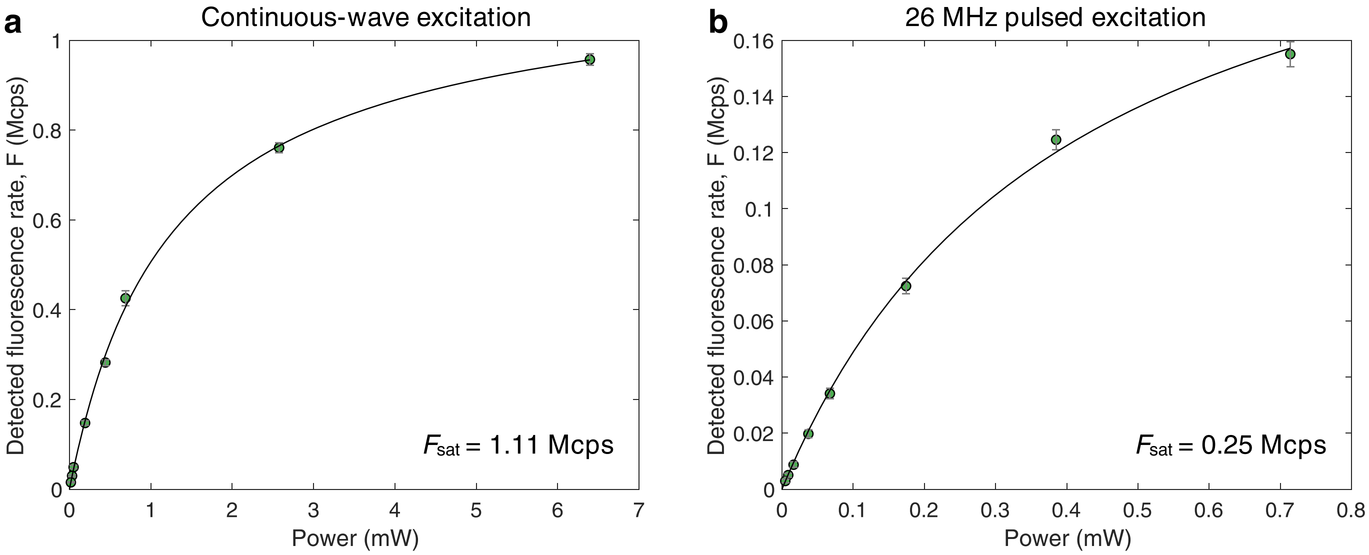
Extended Data Fig. 1 | Flowchart for large-scale heterogeneous integration. See main text and methods for process descriptions.



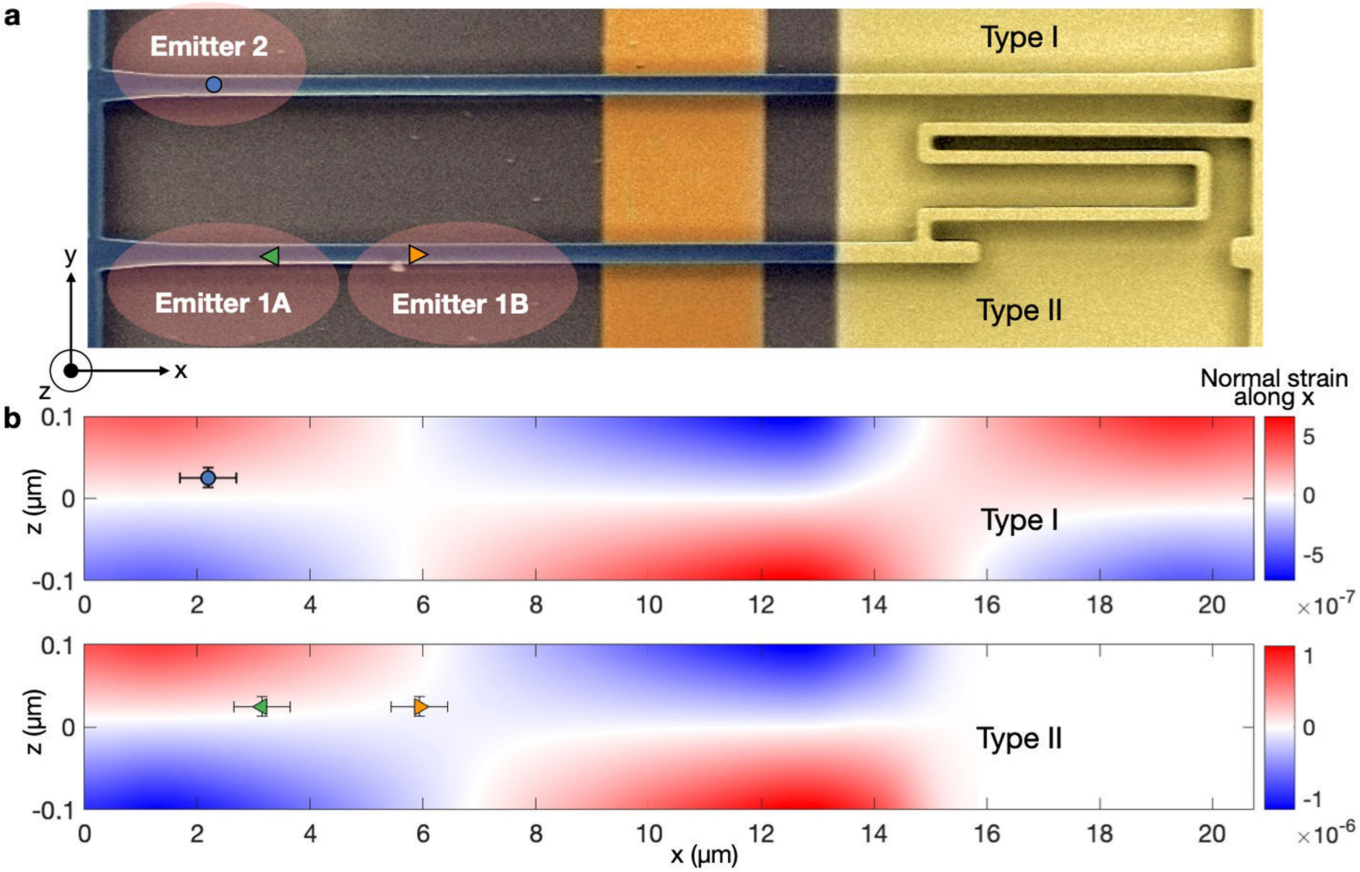
Extended Data Fig. 2 | Histogram of number of emitter-coupled waveguides within a QMC. The red coloured bar corresponds to the defect-free 8-channel QMCs that were suitable for integration. The orange coloured bars correspond to the QMCs that we did not use in this work.



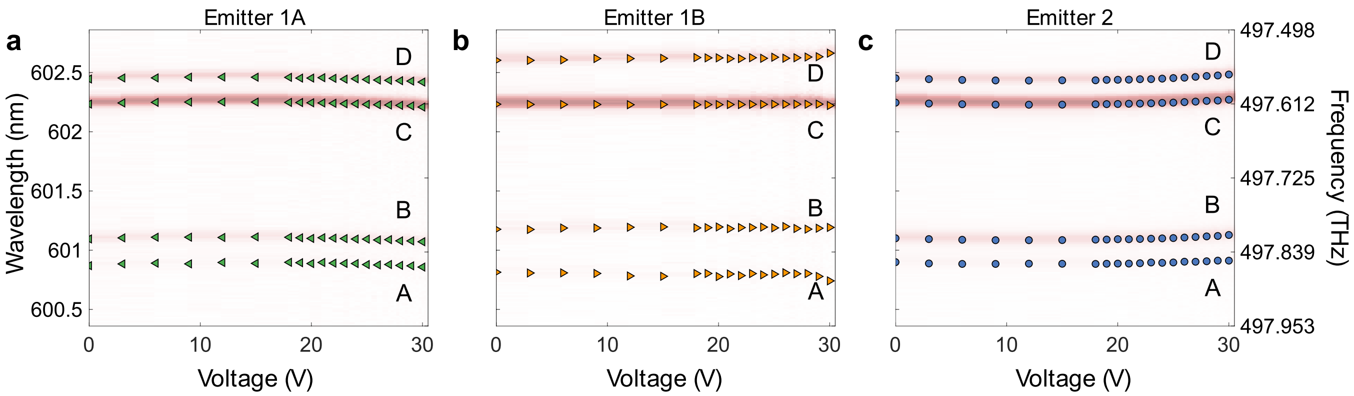
Extended Data Fig. 3 | FDTD simulation showing propagation of light from the diamond waveguide into the AlN waveguide. a, For a 602-nm wavelength (corresponding to the GeV colour centre ZPL). **b**, For a 737-nm wavelength (corresponding to the SiV colour centre ZPL).



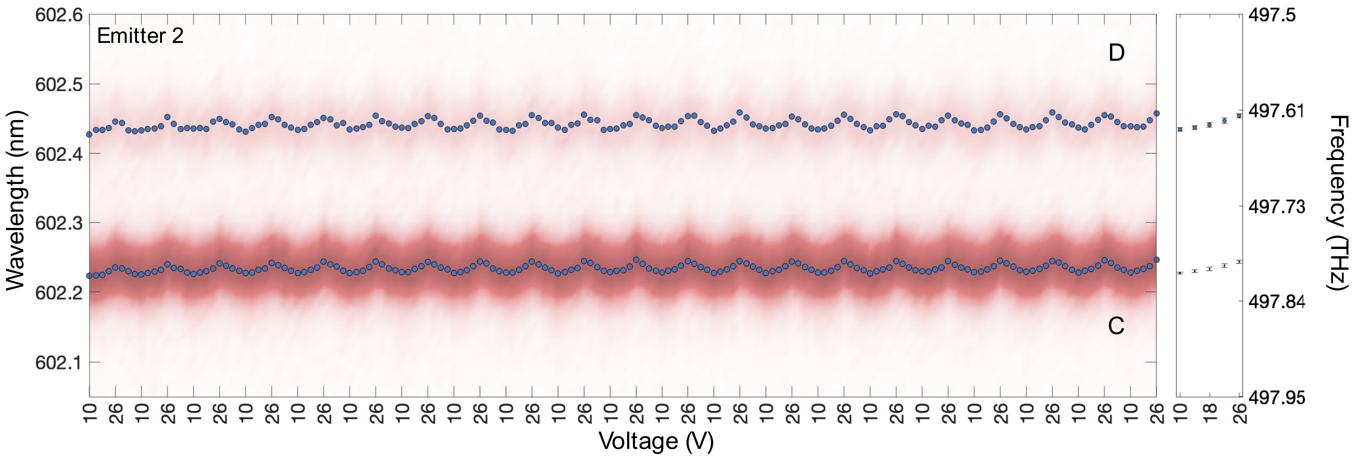
Extended Data Fig. 4 | Saturation response of a single GeV centre. a, Continuous-wave 532-nm laser excitation **b**, Pulsed laser excitation at 532 nm with a repetition rate of 26 MHz.



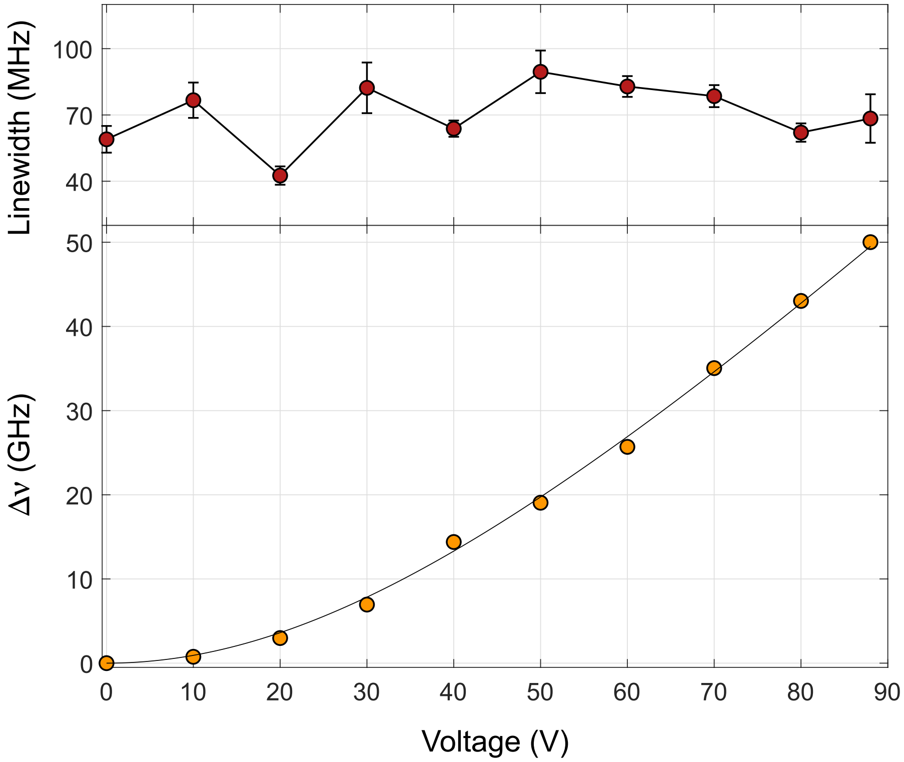
Extended Data Fig. 5 | Scheme for strain-tuning emitters in a PIC platform. **a**, SEM image of type I and type II waveguides considered in this experiment. **b**, Strain distribution along the waveguides and emitters considered in the main text (Fig. 5). Horizontal error bars indicate the lateral uncertainty in the position of emitters and vertical error bars indicate the ion implantation straggle.



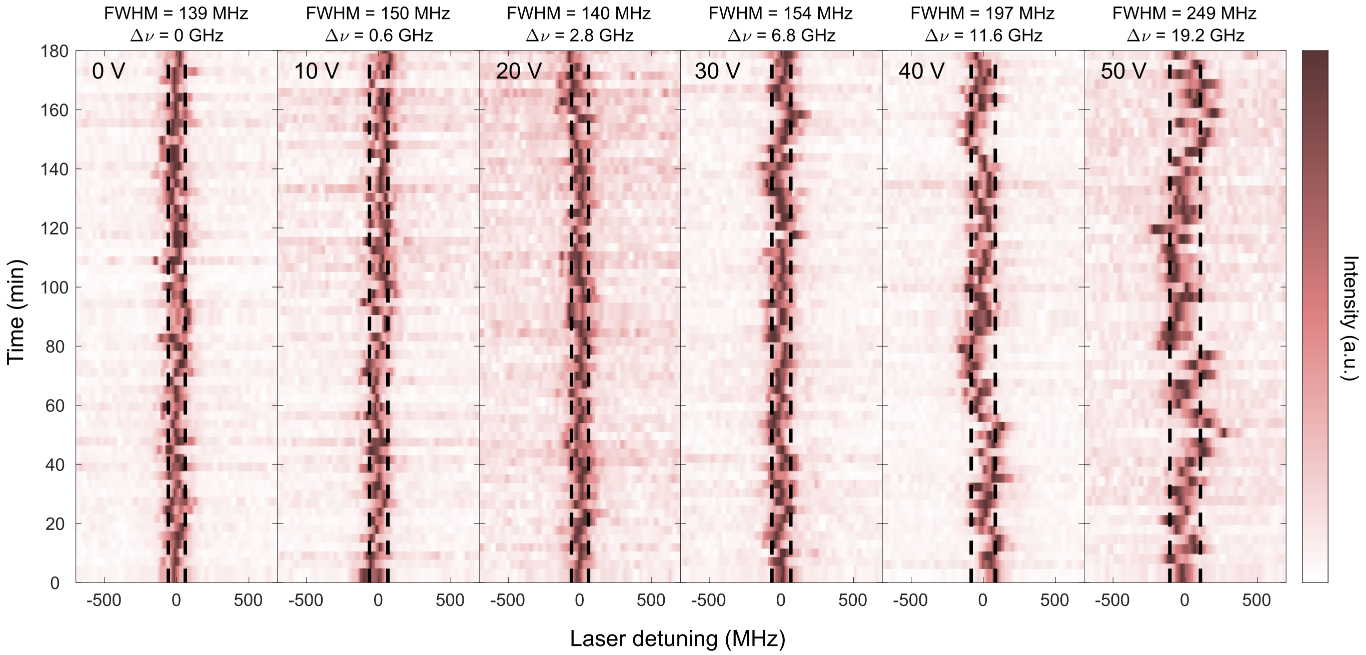
Extended Data Fig. 6 | Spectral shift of GeV centres in response to strain fields. a–c, Strain response of emitter 1A (**a**), emitter 1B (**b**) and emitter 2 (**c**).



Extended Data Fig. 7 | Spectral shifts for the brightest transitions. Reproducible spectral shifts between 10 V and 26 V for the two brightest transitions C and D for emitter 2.



Extended Data Fig. 8 | Optical properties during strain tuning. Top: PLE linewidths as a function of voltage. Bottom: corresponding frequency shift, Δ*ν*, of the ZPL transition.



Extended Data Fig. 9 | Stability of the ZPL transition frequency during strain tuning. Each time slice corresponds to a single PLE linewidth measurement averaged over 2,000 experiments (about 3 min).

Extended Data Table 1 | Saturated count rates from single GeV centres in a QMC

Channel	41	42	43	44	45	46	47	48
Saturated counts x 10^6 cps	0.20(3)	0.94(24)	0.21(7)	0.80(9)	0.61(2)	0.91(13)	1.11(23)	0.31(09)

Coherent control of a surface structural phase transition

https://doi.org/10.1038/s41586-020-2440-4

Received: 21 June 2019

Accepted: 31 March 2020

Jan Gerrit Horstmann[1], Hannes Böckmann[1], Bareld Wit[1], Felix Kurtz[1], Gero Storeck[1] & Claus Ropers[1,2]✉

Active optical control over matter is desirable in many scientific disciplines, with prominent examples in all-optical magnetic switching[1,2], light-induced metastable or exotic phases of solids[3–8] and the coherent control of chemical reactions[9,10]. Typically, these approaches dynamically steer a system towards states or reaction products far from equilibrium. In solids, metal-to-insulator transitions are an important target for optical manipulation, offering ultrafast changes of the electronic[4] and lattice[11–16] properties. The impact of coherences on the efficiencies and thresholds of such transitions, however, remains a largely open subject. Here, we demonstrate coherent control over a metal–insulator structural phase transition in a quasi-one-dimensional solid-state surface system. A femtosecond double-pulse excitation scheme[17–20] is used to switch the system from the insulating to a metastable metallic state, and the corresponding structural changes are monitored by ultrafast low-energy electron diffraction[21,22]. To govern the transition, we harness vibrational coherence in key structural modes connecting both phases, and observe delay-dependent oscillations in the double-pulse switching efficiency. Mode-selective coherent control of solids and surfaces could open new routes to switching chemical and physical functionalities, enabled by metastable and non-equilibrium states.

Femtochemistry entails the search for understanding and control of ultrafast reaction pathways[9,20]. To this end, coherences in the electronic and vibrational states of reactants are used to affect transitions in a complex, generally multidimensional energy landscape[9,23]. Established for small molecules, the possible transfer of this concept to extended systems and solids is complicated by, for example, the high electronic and vibrational density of states, and by couplings to an external heat bath. Low-dimensional and strongly correlated systems represent a promising intermediate between molecules and solids, with phase transitions assuming the role of a 'reaction'. Some of these transitions can be driven optically by means of transient heating[22,24], electronic excitation[13–16,25,26] or direct resonant coupling to certain vibrational degrees of freedom[4–6].

The prototypical case of a phase transition governed by structural modes is given by the Peierls instability[27], in which a metal-to-insulator transition is linked to phonon softening and the appearance of a static periodic lattice distortion. Coherent oscillations of the periodic lattice distortion, known as amplitude modes or amplitudons, are frequently observed in the optical pumping of such transitions, especially close to their thresholds[28–33]. In analogy to the vibrational spectroscopy of reacting molecules[34], amplitudons can be used to track ultrafast changes in the lattice symmetry across a phase transition[29,32,33]. However, it remains to be shown how coherent amplitude motion can be used to manipulate the outcome of a structural transition.

Here we report coherent control over the phase transition in a quasi-one-dimensional Peierls insulator by means of the amplitudes of specific phonon modes. We use a double-pulse excitation scheme and monitor the structural transformation by ultrafast low-energy electron diffraction (ULEED; Fig. 1a, see Methods)[21,22]. Observing the resulting structure as a function of the double-pulse separation demonstrates the importance of shear and rotational phonon modes on the femtosecond timescale. A comparison of ULEED and transient reflectivity measurements suggests distinct roles of these phonons in controlling the transition, and points to the location of the transition state along the mode coordinates.

As a model system, we study atomic indium wires on the (111) surface of silicon[35], a prominent Peierls system attracting interest for its ultrafast dynamics[13–16]. Arranged in a zigzag pattern, the indium atoms induce a metallic (4 × 1) superstructure, which, at critical temperature $T_c = 125$ K, exhibits a first-order transition to an insulating state with quadrupled (8 × 2) unit cell size and a hexagon-shaped indium pattern. The associated change in atomic structure causes additional spots in backscattering diffraction (see LEED patterns in Fig. 1b). Below T_c, a single optical pump pulse is able to electronically excite the system to a metastable (4 × 1) state[13–16]. Time-resolved diffraction and photoemission spectroscopy have recently revealed the ultrafast and ballistic nature of this transition (occurring on a 350-fs timescale) and identified excited electrons and localized photoholes as its driving force[14–16].

Tracking the (4 × 1) and (8 × 2) diffraction spot intensities in ULEED, we observe a rapid increase/decrease directly after optical excitation and subsequent relaxation to a level persisting over nanoseconds (Fig. 1c, left), indicating the metastability of the structure[13]. Interestingly, this long-lived contribution displays a rather gradual threshold in pump fluence. This implies that for intermediate excitation

---

[1]4th Physical Institute, Solids and Nanostructures, University of Göttingen, Göttingen, Germany. [2]Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. ✉e-mail: cropers@gwdg.de

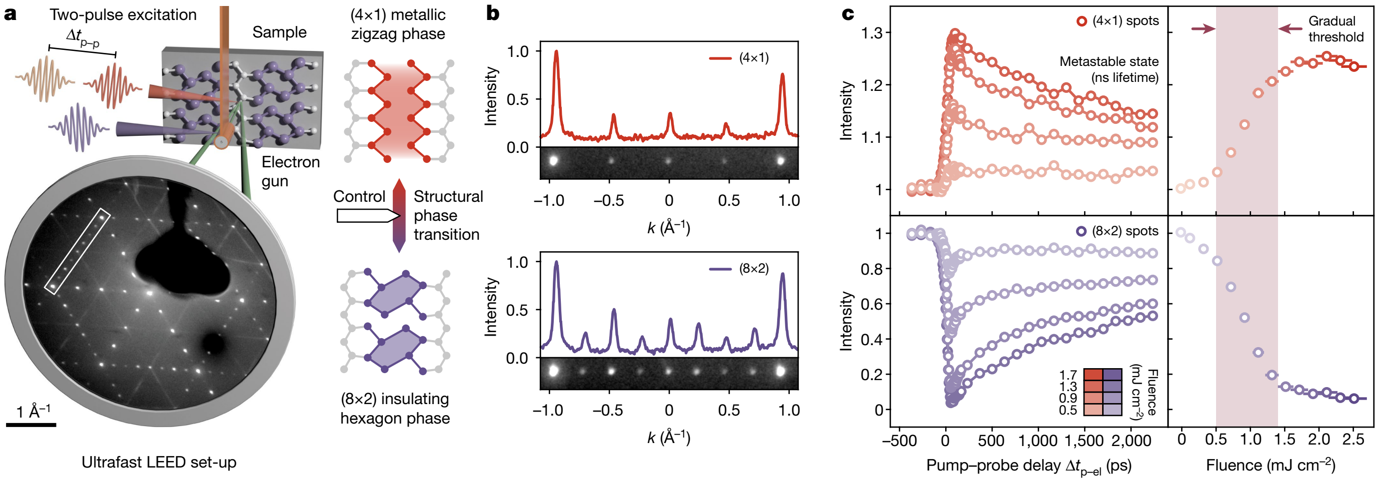


Fig. 1 | Ultrafast LEED set-up and structural phase transition in atomic indium wires on silicon. a, Experimental scheme. Ultrashort electron pulses from a miniaturized laser-driven electron gun are used in a LEED experiment to monitor the microscopic structure of atomic indium wires on the Si(111) surface after optical excitation by single or double pulses. White, silicon atoms; violet, indium. **b**, Selected regions and line profiles from LEED patterns of the metallic (4 × 1) and insulating (8 × 2) phases (white frame in **a**). The emergence of additional spots in the (8 × 2) phase indicates the pronounced structural changes during the phase transition. **c**, Left, time-resolved integrated intensities of (4 × 1) and (8 × 2) diffraction spots as a function of the pump–probe delay $\Delta t_{p\text{-}el}$. Right, fluence-dependent spot intensities recorded at $\Delta t_{p\text{-}el} = 75$ ps. The (4 × 1) and (8 × 2) intensities have been normalized to corresponding values at $\Delta t_{p\text{-}el} < 0$. Errorbars, experimental uncertainty in fluence determination.

densities, a variable part of the surface is switched to the metastable state (Fig. 1c, right), despite homogeneous excitation of the probed area (see Methods). An interpretation based on the coexistence of both phases is also corroborated by scanning tunnelling microscopy[36] and Raman spectroscopy[37] well below T_c.

It seems likely that near the threshold, the structural transition is particularly susceptible to weak perturbations, affecting the efficiency of driving the system to the metastable state. Motivated by control schemes in femtochemistry[17–20], we explore the use of pulse sequences to manipulate the switching efficiency. Specifically, we use a pair of optical pump pulses with variable delay $\Delta t_{p\text{-}p}$ and probe the resulting structure by ULEED at a later time of $\Delta t_{p\text{-}el} = 75$ ps, well after the excitation. We find that the signature of the metastable state—that is, a suppression of the (8 × 2) and increase of the (4 × 1) phase—is a strong function of the double-pulse delay $\Delta t_{p\text{-}p}$ (Fig. 2a). Importantly, at intermediate fluences between 0.5 mJ cm^{-2} and 1.4 mJ cm^{-2}, pronounced oscillations with a period of 1–2 ps are observed on either delay side, with opposing behaviour for the (4 × 1) (top panel) and (8 × 2) spots. By contrast, only a minor delay-dependence is found well below and above threshold. The peaked signal around $\Delta t_{p\text{-}p} = 0$ is attributed to additive electronic excitation which decays on a few-picosecond timescale[38].

The observed oscillations clearly demonstrate a coherent response of the signal. In particular, coherent vibrational motion induced by the first pulse controls the switching efficiency for the second pulse. The frequency content of the signal (Fig. 2b, top) points to shear and rotation phonon modes, which have previously been identified as amplitude modes of the metal–insulator phase transition[37,39] (see Fig. 2c). Interestingly, Fourier-filtered traces of the two observed frequency bands exhibit opposite phases at time zero, corresponding to enhancement or suppression of the transition for pulse overlap (see Fig. 2b, bottom).

The appearance of coherent phonons is understood within the established potential energy model of the transition[13,14,39,40]. The (8 × 2)→(4 × 1) transformation is typically described in terms of a tristable energy surface, with an initial minimum at the (8 × 2) configuration. Electronic excitation tilts the balance towards the (4 × 1) phase (Fig. 3a)[13–15,40], accompanied by displacive excitation of coherent phonons (DECP)[25]. The question now arises as to how these phonon coherences enable control over the transition.

Generally, Raman-active phonons can modulate the optical absorption of a surface[25,29], affecting the level of electronic excitation achieved by the second pulse. This can influence the final-state potential energy surface and the observed threshold ('absorption control'; see Fig. 3c). Amplitude modes of the symmetry-broken state are expected to play a role in this mechanism, given their direct link to the structural transformation and their susceptibility to strong displacive excitation. Moreover, the ballistic nature of the transition[14] and the influence of DECP suggest that kinetic energy contributes to overcoming a sufficiently lowered but not completely vanishing barrier ('ballistic control'; see Fig. 3b). For the vibrational motion along a reaction coordinate, in-phase excitation with a second pulse maximizes the effect of DECP and allows barrier-crossing to the (4 × 1) state (1). Anti-phase excitation, on the other hand, vibrationally de-excites the system, which then has insufficient kinetic energy and remains in the (8 × 2) state (2). In a corresponding real-space picture, by weakening or strengthening different indium–indium bonds[15] and thus shifting the equilibrium atomic positions, the second pulse either adds further mechanical stress to the system (1) or removes it (2). Whereas the absorption modulation described above may apply to all Raman-active modes q (ref. [25]), this ballistic contribution is only feasible for modes along the reaction coordinate Q.

To further elucidate the contributions of these mechanisms, we complement ULEED by optical pump–probe (OPP) spectroscopy (see Methods for details), which probes absorption modulation by coherent phonons[25]. We measure pump-induced changes in the optical reflectivity, which are directly proportional to absorption changes for a monolayer on a substrate with real refractive index[41]. A comparison of the switching efficiency (Fig. 3d) with the transient reflectivity (Fig. 3e) reveals both similarities and stark differences in the observed frequencies and their relative amplitudes (see also Extended Data Fig. 6 for further OPP traces).

Both measurements exhibit a frequency component close to 0.82 THz, which we assign to the hexagon rotation mode of the (8 × 2) structure[37,39]. Owing to its anti-phase behaviour at $\Delta t_{p\text{-}p} = 0$ (see Fig. 2b), for this mode, a ballistic control mechanism can be ruled out. We therefore attribute the rotation-mode ULEED signal to absorption modulation. This assignment is further corroborated by the modulation amplitudes in both types of measurements, which are linked through

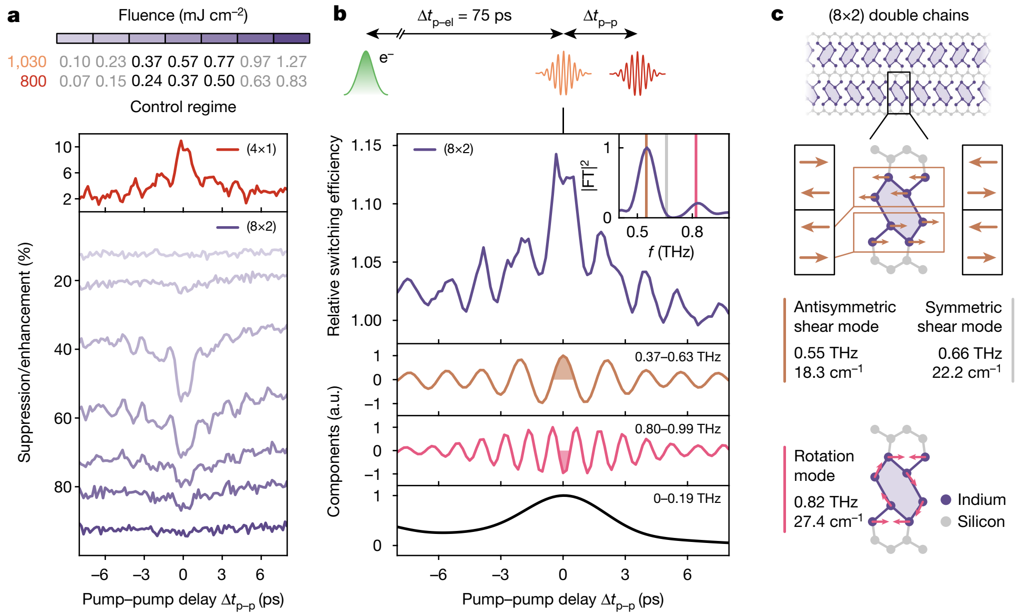


Fig. 2 | Coherent control of the (8 × 2) to (4 × 1) phase transition efficiency. **a**, Suppression or enhancement of the integrated (8 × 2) and (4 × 1) diffraction spot intensity as a function of the double-pulse delay Δt_{p-p} and incident fluence F. (4 × 1) trace: $F_{1,030} = 0.37$ mJ cm^{-2}, $F_{800} = 0.24$ mJ cm^{-2}. **b**, Top, delay-dependent relative switching efficiency for $F_{1,030} = 0.32$ mJ cm^{-2}, $F_{800} = 0.21$ mJ cm^{-2}. Inset, spectral density of switching efficiency; vertical lines represent the frequencies of structural modes given in **c**. FT, Fourier transform. Bottom, Fourier-filtered contributions of different frequency components. Brown and pink shaded regions indicate the distinct initial phases. a.u., arbitrary units. **c**, Prominent low-frequency modes of the (8 × 2) structure.

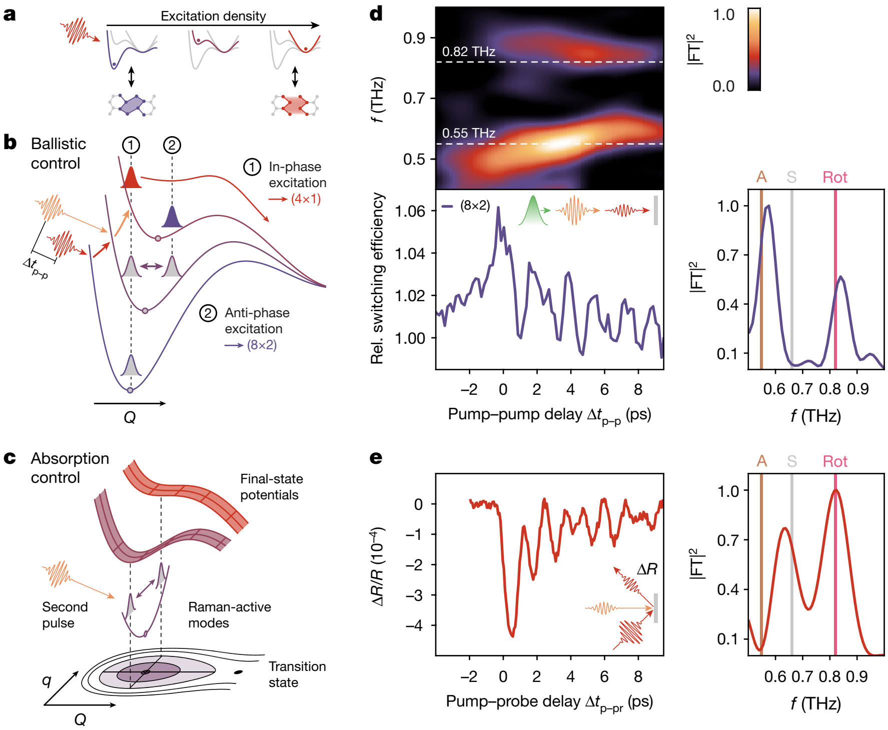


Fig. 3 | Control mechanisms and comparison between ULEED and optical pump–probe spectroscopy. **a**, Phase transition model based on reshaping of the tristable energy surface by a single pump pulse. For simplicity, the second, energetically degenerate (8 × 2) minimum is not depicted. Note that the potential deformation is a continuous function of the excitation density. **b**, **c**, Coherent control mechanisms in double-pulse experiments: ballistic control (**b**) and absorption control (**c**). **d**, Relative switching efficiency recorded for unequal pump pulses in ULEED ($F_{1,030} = 0.48$ mJ cm^{-2}, $F_{800} = 0.15$ mJ cm^{-2}), corresponding spectral density (right) with reference frequencies (see Fig. 2c) and short-time Fourier transform (top). A, S and Rot indicate the frequencies of antisymmetric, symmetric and rotation modes, respectively. **e**, Delay-dependent reflectivity changes $\Delta R/R$ of the surface measured in optical pump–probe experiments and corresponding spectral density ($F_{\text{pump}} = 0.15$ mJ cm^{-2}). Δt_{p-pr}, delay between optical pump and optical probe pulses.

the total absorption of the monolayer. Our measurements predict a value of about 1% for the absorption, similar to a recent estimate[16] (see Methods).

A more intricate situation is found for the low-frequency component, associated with shear phonons: this dominant feature in ULEED modulates the switching efficiency to a disproportionately higher degree than expected from the overall transient reflectivity. Moreover, the shear mode frequencies measured by ULEED (0.57 THz) and OPP (0.64 THz) differ significantly. This may be a result of OPP probing a surface-averaged optical response, whereas ULEED is sensitive to the transition probability in regions close to threshold. However, density functional theory and Raman spectroscopy in fact predict two separate shear modes: whereas the symmetric shear mode (expected at 0.66 THz) is much more prominent in Raman spectra[37], only the antisymmetric shear mode (0.55 THz) is considered relevant for the transition[42,43]. These distinct properties suggest that each of OPP and ULEED mainly probes a different one of these modes, namely the higher-frequency symmetric and the lower-frequency antisymmetric shear oscillation, respectively.

From these considerations, we extract two possible scenarios for the role of shear motion, linked to the control mechanisms discussed above (Fig. 3). First, if the transition is indeed driven by a shear mode separate from that seen in reflectivity, we must invoke the ballistic mechanism (Fig. 3b) to explain the ULEED data, directly linking this mode to the reaction coordinate. Alternatively, to identify the shear contributions in ULEED and OPP with the same mode and absorption modulation (Fig. 3c), the observed frequency difference requires further explanation. In particular, this would necessitate a greatly softened and larger-amplitude shear mode oscillation only in surface regions that can be switched by the second pulse, with an unaltered rotation frequency (Fig. 3d).

Both scenarios imply that the shear displacement corresponds to the primary reaction coordinate, whereas the rotation completing the transition[39,42] is of a secondary nature. Accordingly, we propose a description of the transition in terms of a two-dimensional potential-energy surface spanned by the rotation and shear deformations of the (4 × 1) structure (Fig. 4a), with the system initially residing in the (8 × 2) minimum. In a reasonable assumption, the first pulse induces a displacive excitation of coherent phonons towards the (4 × 1) state. The ULEED measurements show that the transition efficiency for the second pulse becomes a strong function of the momentary vibrational state (Fig. 4b), denoted by the colour-coded area in Fig. 4b. The combined observations—that is, the differences in frequency and relative amplitudes between ULEED and OPP, as well as the phases of both modes in the double-pulse traces (Fig. 2b)—now suggest an 'off-diagonal' transition state in configuration space with a strongly reduced shear but a largely unaltered rotation (compared with the (8 × 2) state). This interpretation is further supported by the transient softening and hardening (Fig. 3d) of the shear and rotation mode, respectively, near $\Delta t_{\text{p-p}} = 0$ (Fig. 3d), which we have consistently observed in a number of experiments (see Extended Data Fig. 7).

In the proposed pathway of the transition, overcoming an 'early' barrier[23], the In-chains are first 'un-sheared' and subsequently transformed into the zigzag structure by a rotation. Such a pathway transiently passes the 'trimer' state[43,44], which has been intensely studied by density functional theory and is expected to be almost energetically degenerate to the (4 × 1) state[45]. The existence of two local minima for a similar rotation displacement, namely the trimer and the (8 × 2) configuration[45], also supports the existence of a transition state along the shear axis.

Future experimental and theoretical studies, involving density functional theory and molecular dynamics simulations, may further elucidate the potential energy surface, possible additional pathways and the sequential nature of the transition. The microscopic excitation mechanism underlying the phonon coherences deserves further consideration, including its link to the femtosecond electron transfer and hole-induced driving forces recently described[15]. Finally, considering the surface heterogeneity, the influence of frequency changes at domain boundaries[36] on the local transition dynamics will be of interest.

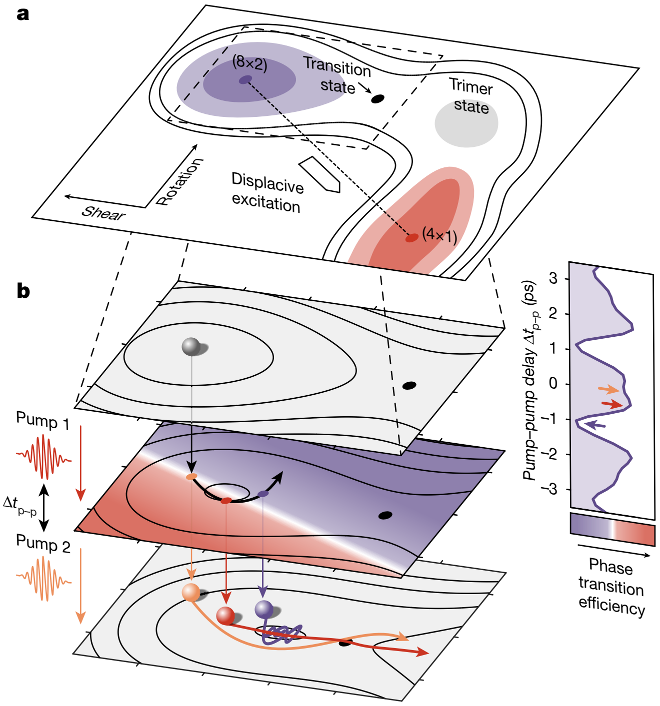


Fig. 4 | Two-dimensional picture of the phase transition dynamics. **a**, Proposed two-dimensional model of the potential energy surface for the (8 × 2)→(4 × 1) in shear/rotation configuration space, exhibiting a transition state along the shear axis from the (8 × 2) state. **b**, Sketch of exemplary system trajectories close to the (8 × 2) state before (top), in between (middle) and after (bottom) two subsequent displacive excitations (yellow, $\Delta t_{\text{p-p}} = 0$; red, $\Delta t_{\text{p-p}} \approx T_{\text{rot}}/2$; violet, $\Delta t_{\text{p-p}} \approx T_{\text{shear}}/2$). T_{shear} and T_{rot}, oscillation period of the shear and rotation mode, respectively. The phase transition efficiency (colour-coded) is a strong function of the vibrational coordinates at the time of the second pulse (middle and right panel). Highest efficiency is achieved for a maximum sheared/minimum rotated structure (see middle panel).

Our results demonstrate the coherent control of a surface structural phase transition by all-optical manipulation of key phonon modes, and show that the outcome of the phase transition, much like many chemical reactions, depends on the momentary state of the coherent vibrational wavepacket. Close to the transition threshold, both absorption modulation by Raman-active modes and the ballistic motion of the order parameter in overcoming the barrier should be considered. The latter contribution could be enhanced with mode-selectivity by a repeated stimulation of the coherent phonon amplitude, which, as in the present system, decays slower than the electronic excitation[18,46].

In molecular chemistry, it has long been known that vibrational excitation and the location of the transition state may greatly affect reaction rates, a principle captured by the Polanyi rules[23]. Our work extends this principle to surfaces and solids and introduces the vibrational phase as a decisive parameter to target the transition state. We believe that exploiting vibrational coherences in low-dimensional and strongly correlated materials, as well as molecular adsorbates, holds promise for structural and electronic control in surface physics and chemistry, providing a handle to steer physical functionality and chemical reactivity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2440-4.

1. Kimel, A. V. et al. Ultrafast non-thermal control of magnetization by instantaneous photomagnetic pulses. *Nature* **435**, 655–657 (2005).
2. Schlauderer, S. et al. Temporal and spectral fingerprints of ultrafast all-coherent spin switching. *Nature* **569**, 383–387 (2019).
3. Stojchevska, L. et al. Ultrafast switching to a stable hidden quantum state in an electronic crystal. *Science* **344**, 177–180 (2014).
4. Rini, M. et al. Control of the electronic phase of a manganite by mode-selective vibrational excitation. *Nature* **449**, 72–74 (2007).
5. Mitrano, M. et al. Possible light-induced superconductivity in K_3C_{60} at high temperature. *Nature* **530**, 461–464 (2016).
6. Nova, T. F., Disa, A. S., Fechner, M. & Cavalleri, A. Metastable ferroelectricity in optically strained $SrTiO_3$. *Science* **364**, 1075–1079 (2019).
7. Sie, E. J. et al. An ultrafast symmetry switch in a Weyl semimetal. *Nature* **565**, 61–66 (2019).
8. Wang, Y. H., Steinberg, H., Jarillo-Herrero, P. & Gedik, N. Observation of Floquet–Bloch states on the surface of a topological insulator. *Science* **342**, 453–457 (2013).
9. Zewail, A. H. Femtochemistry: atomic-scale dynamics of the chemical bond using ultrafast lasers (Nobel lecture). *Angew. Chem. Int. Ed*. **39**, 2586–2631 (2000).
10. Nuernberger, P., Vogt, G., Brixner, T. & Gerber, G. Femtosecond quantum control of molecular dynamics in the condensed phase. *Phys. Chem. Chem. Phys*. **9**, 2470–2497 (2007).
11. Morrison, V. R. et al. A photoinduced metal-like phase of monoclinic VO_2 revealed by ultrafast electron diffraction. *Science* **346**, 445–448 (2014).
12. Liu, M. et al. Terahertz-field-induced insulator-to-metal transition in vanadium dioxide metamaterial. *Nature* **487**, 345–348 (2012).
13. Wall, S. et al. Atomistic picture of charge density wave formation at surfaces. *Phys. Rev. Lett*. **109**, 186101 (2012).
14. Frigge, T. et al. Optically excited structural transition in atomic wires on surfaces at the quantum limit. *Nature* **544**, 207–211 (2017).
15. Nicholson, C. W. et al. Beyond the molecular movie: dynamics of bands and bonds during a photoinduced phase transition. *Science* **362**, 821–825 (2018).
16. Chávez-Cervantes, M., Krause, R., Aeschlimann, S. & Gierz, I. Band structure dynamics in indium wires. *Phys. Rev. B* **97**, 201401 (2018).
17. Hase, M., Fons, P., Mitrofanov, K., Kolobov, A. V. & Tominaga, J. Femtosecond structural transformation of phase-change materials far from equilibrium monitored by coherent phonons. *Nat. Commun*. **6**, 8367 (2015).
18. Weiner, A. M., Leaird, D. E., Wiederrecht, G. P. & Nelson, K. A. Femtosecond pulse sequences used for optical manipulation of molecular motion. *Science* **247**, 1317–1319 (1990).
19. Feurer, T., Vaughan, J. C. & Nelson, K. A. Spatiotemporal coherent control of lattice vibrational waves. *Science* **299**, 374–377 (2003).
20. Potter, E. D., Herek, J. L., Pedersen, S., Liu, Q. & Zewail, A. H. Femtosecond laser control of a chemical reaction. *Nature* **355**, 66 (1992).
21. Gulde, M. et al. Ultrafast low-energy electron diffraction in transmission resolves polymer/graphene superstructure dynamics. *Science* **345**, 200–204 (2014).
22. Vogelgesang, S. et al. Phase ordering of charge density waves traced by ultrafast low-energy electron diffraction. *Nat. Phys*. **14**, 184–190 (2018).
23. Polanyi, J. C., Wong, W. H. & Mok, M. H. Location of energy barriers. *J. Chem. Phys*. **51**, 1439–1450 (1969).
24. Haupt, K. et al. Ultrafast metamorphosis of a complex charge-density wave. *Phys. Rev. Lett*. **116**, 016402 (2016).
25. Zeiger, H. J. et al. Theory for displacive excitation of coherent phonons. *Phys. Rev. B* **45**, 768–778 (1992).
26. Sciaini, G. et al. Electronic acceleration of atomic motions and disordering in bismuth. *Nature* **458**, 56–59 (2009).
27. Peierls, R. E. *Quantum Theory of Solids* (Oxford Univ. Press, 2001).
28. Eichberger, M. et al. Snapshots of cooperative atomic motions in the optical suppression of charge density waves. *Nature* **468**, 799–802 (2010).
29. Wall, S. et al. Ultrafast changes in lattice symmetry probed by coherent phonons. *Nat. Commun*. **3**, 721 (2012).
30. Sokolowski-Tinten, K. et al. Femtosecond X-ray measurement of coherent lattice vibrations near the Lindemann stability limit. *Nature* **422**, 287–289 (2003).
31. Rettig, L., Chu, J.-H., Fisher, I. R., Bovensiepen, U. & Wolf, M. Coherent dynamics of the charge density wave gap in tritellurides. *Faraday Discuss*. **171**, 299–310 (2014).
32. Beaud, P. et al. A time-dependent order parameter for ultrafast photoinduced phase transitions. *Nat. Mater*. **13**, 923–927 (2014).
33. Neugebauer, M. J. et al. Optical control of vibrational coherence triggered by an ultrafast phase transition. *Phys. Rev. B* **99**, 220302 (2019).
34. Nibbering, E. T. J., Fidder, H. & Pines, E. Ultrafast chemistry: using time-resolved vibrational spectroscopy for interrogation of structural dynamics. *Annu. Rev. Phys. Chem*. **56**, 337–367 (2005).
35. Yeom, H. W. et al. Instability and charge density wave of metallic quantum chains on a silicon surface. *Phys. Rev. Lett*. **82**, 4898–4901 (1999).
36. Song, S. K., Samad, A., Wippermann, S. & Yeom, H. W. Dynamical metal to charge-density-wave junctions in an atomic wire array. *Nano Lett*. **19**, 5769–5773 (2019).
37. Speiser, E., Esser, N., Wippermann, S. & Schmidt, W. G. Surface vibrational Raman modes of In:Si(111) (4 × 1) and (8 × 2) nanowires. *Phys. Rev. B* **94**, 075417 (2016).
38. Nicholson, C. W. et al. Excited-state band mapping and momentum-resolved ultrafast population dynamics in In/Si(111) nanowires investigated with XUV-based time- and angle-resolved photoemission spectroscopy. *Phys. Rev. B* **99**, 155107 (2019).
39. Wippermann, S. & Schmidt, W. G. Entropy explains metal–insulator transition of the Si(111)–In nanowire array. *Phys. Rev. Lett*. **105**, 126102 (2010).
40. Jeckelmann, E., Sanna, S., Schmidt, W. G., Speiser, E. & Esser, N. Grand canonical Peierls transition in In/Si(111). *Phys. Rev. B* **93**, 241407 (2016).
41. Li, Y. & Heinz, T. F. Optical models for thin layers. Preprint at https://arxiv.org/abs/1801.00402 (2018).
42. Wippermann, S. Understanding Substrate-Supported Atomic-Scale Nanowires from Ab Initio Theory. PhD thesis, Univ. Paderborn (2010).
43. Stekolnikov, A. A. et al. Hexagon versus trimer formation in In nanowires on Si(111): energetics and quantum conductance. *Phys. Rev. Lett*. **98**, 026105 (2007).
44. Kumpf, C. et al. Low-temperature structure of indium quantum chains on silicon. *Phys. Rev. Lett*. **85**, 4916–4919 (2000).
45. González, C., Ortega, J. & Flores, F. Metal–insulator transition in one-dimensional In-chains on Si(111): combination of a soft shear distortion and a double-band Peierls instability. *New J. Phys*. **7**, 100 (2005).
46. Nelson, K. A. in *Mode Selective Chemistry* (eds Jortner, J. et al.) 527–533 (Springer, 1991).

Methods

Ultrafast LEED set-up

We recently developed ULEED in an optical-pump/electron-probe scheme for the time-resolved investigation of structural dynamics at solid-state surfaces[21,22,47]. LEED is a surface-sensitive technique, in which the diffraction pattern of electrons backscattered from a sample is analysed to obtain information about the surface structure[48].

To achieve high temporal and momentum resolution, we use a laser-driven electron gun consisting of a nanometric tungsten tip as well as four metal electrodes (outer diameter 2 mm, aperture diameter 400 μm), which act as a suppressor–extractor unit and an electrostatic einzel lens[22]. Electron pulses are generated by localized two-photon photoemission by illuminating the tip apex with femtosecond laser pulses (central wavelength 400 nm, pulse duration 45 fs, pulse energy 20 nJ) at repetition rates up to 100 kHz (note that the data presented in Figs. 2b and 3d were recorded at a repetition rate of 25 kHz whereas all other ULEED measurements were carried out at a repetition rate of 100 kHz). The needle cathode provides a reduced electron beam emittance, allowing for a momentum resolution in diffraction of 0.03 Å^{-1}. Moreover, we lower the dispersion-induced broadening effect on the electron pulse by decreasing the propagation length between the electron source and the sample. In this respect, the reduced dimensions of the electron gun allow for operational distances of a few millimetres at a reasonably small fraction of shadowed electron diffraction signal, resulting in electron-pulse durations down to 16 ps (depending on gun–sample distance)[22]. The backscattered electrons from the surface are amplified and recorded by a combination of a chevron microchannel plate, a phosphor screen and a cooled sCMOS (scientific complementary metal–oxide–semiconductor) camera resulting in typical integration times of $t_{\mathrm{int}} = 20$ s per frame in time-resolved measurements.

In ULEED pump–probe experiments (Fig. 1c), the surface structure is excited by ultrashort light pulses ($\lambda_c = 1{,}030$ nm, $\hbar\omega_1 = 1.2$ eV, $\Delta\tau = 212$ fs) from an Yb:YAG amplifier system and probed by electron pulses (kinetic energy $E_{\mathrm{kin}} = 80$ eV) at a variable time delay $\Delta t_{\mathrm{p-el}}$. To ensure a homogeneous excitation of the area probed by the electrons, we expand the optical pump beam to (297 ± 13) μm × (223 ± 14) μm in the sample plane, which is considerably larger than the focal spot size of the electron gun (<80 μm × 80 μm). The electron beam diameter corresponds to at least hundreds of structural correlation lengths (taken from scanning tunnelling microscopy literature; see, for example, refs. [36,49,50]), thus averaging over a large ensemble of local configurations.

For the coherent control of the structural phase transition between the (4 × 1) and the (8 × 2) phase (Fig. 2a,b; Fig. 3d), we use two pump pulses P_1 and P_2 with distinct central wavelengths λ_c (P_1: $\lambda_c = 1{,}030$ nm, $\hbar\omega_1 = 1.2$ eV, $\Delta\tau = 212$ fs; P_2: $\lambda_c = 800$ nm, $\hbar\omega_2 = 1.55$ eV, $\Delta\tau = 232$ fs) from the amplifier system and an optical parametric amplifier (OPA) to avoid interference effects around time-zero (coherent artefacts). The P_1 and P_2 beams are aligned collinearly and subsequently focused onto the sample by a single lens (see Extended Data Fig. 1a). To determine the temporal overlap of the pump pulses, we perform cross-correlation measurements using a fast nonlinear photodiode (GaP) (see Extended Data Fig. 1b). A sketch of the experimental set-up is depicted in Extended Data Fig. 1a.

Optical pump–probe set-up

To investigate the optical absorption modulation caused by structural modes of the indium monolayer, we use an optical pump–probe set-up to measure the transient reflectivity of the In/Si(111) surface (see Extended Data Fig. 5). In this, a pump pulse ($\lambda_c = 1{,}030$ nm, $f_{\mathrm{rep}} = 100$ kHz) induces coherent phonon oscillations, and the resulting reflectivity changes are monitored by a probe pulse ($\lambda_c = 800$ nm, $f_{\mathrm{rep}} = 100$ kHz) as a function of the time-delay $\Delta t_{\mathrm{p-pr}}$. The pump intensity is modulated at a frequency $f_{\mathrm{mod}} = 25$ kHz by an acousto-optic modulator synchronized to the laser system. Pump and probe pulses are collinearly focused on the sample at an incident angle $\alpha = 31°$. The reflected beam is guided through two short-pass filters (2 × OD4 for $\lambda > 900$ nm) and focused onto a silicon photodiode. The photodiode and reference signals are processed in a lock-in amplifier, yielding the data presented in Fig. 3e and Extended Data Fig. 6.

Sample preparation

All experiments were carried out under ultra-high-vacuum conditions (base pressure $p < 2 \times 10^{-10}$ mbar) to minimize surface defects from adsorption, which were found to have an influence on the formation of the low-temperature (8 × 2) phase as well as the lifetime of the metastable state[13,51]. The samples were prepared by flash-annealing Si(111) wafers (phosphorus-doped, resistivity $R = 0.6$–$2\,\Omega$ cm) at $T_{\mathrm{max}} = 1{,}250$ °C through direct current heating (maximum pressure during flashing was kept below $p_{\mathrm{max}} = 2 \times 10^{-9}$ mbar). Evaporation of 1.2 monolayers of indium onto the resulting Si(111)(7 × 7) surface reconstruction at room temperature followed by subsequent annealing at $T = 500$ °C for 300 s resulted in a high-quality (4 × 1) phase, as verified in our ultrafast LEED set-up. After inspection of the (4 × 1) phase, the samples were immediately cooled to a base temperature of $T = 60$ K with an integrated continuous-flow helium cryostat. The phase transition between the high-temperature (4 × 1) and the low-temperature (8 × 2) phase was observed at 125 K. LEED images of the (7 × 7), the (4 × 1) and the (8 × 2) structures are shown in Extended Data Fig. 2.

Data analysis

The LEED pattern of the (8 × 2) phase from Fig. 1a and the cut-outs (that is, sections of the pattern) shown in Fig. 1b were recorded at a base temperature of $T = 60$ K (cut-out of the (4 × 1) phase: $T = 300$ K) with an integration time of $t_{\mathrm{int}} = 60$ s. The diffraction images are plotted on a logarithmic colour scale to enhance the visibility of the twofold streaks, which are typically one order of magnitude weaker than the (8 × 2) spots. The location of the cut-out regions within the complete diffraction image is indicated by the white rectangle in Fig. 1a.

For the analysis of single- and double-pump ULEED experiments, we sum up the background-corrected raw data peak intensities within circular areas of interest (radius r) around the selected (4 × 1) and (8 × 2) spots. To this end, the background is determined within a ring (width dr) around the edge of each area of interest. We use radii of $r = 0.10$ Å^{-1} (40 pixels) for the fluence-dependent data presented in Fig. 2a, $r = 0.08$ Å^{-1} (30 pixels) for the data presented in Figs. 2b and 3d and a ring width of d$r = 0.008$ Å^{-1} (three pixels) for all datasets. The indices of the analysed spots are listed in Extended Data Fig. 4b.

To determine the relative changes in the (4 × 1) and (8 × 2) spot intensities caused by a single optical pulse (see Fig. 1c), the integrated peak intensities for a saturated suppression/enhancement ($\Delta t_{\mathrm{p-el}} = 75$ ps) are normalized to the value before time-zero. This delay was chosen to account for the finite electron-pulse duration under the conditions of the experiment ($\Delta t_{\mathrm{p-el}} \approx 50$ ps). We consider potential contributions of cumulative heating effects by recording the intensities of both (4 × 1) and (8 × 2) diffraction spots as a function of the sample base temperature T_b (see Extended Data Fig. 3a). For the highest relevant value of fluence ($F \approx 1.35$ mJ cm^{-2}), we find a moderate increase of T_b to a maximum temperature of 82 K, which is well below T_c.

The fluence-dependent enhancement/suppression of the (4 × 1)/(8 × 2) signal in the double-pulse experiments (see Fig. 2a) is shown relative to the intensity $I(\Delta t_{\mathrm{p-el}} = 75\ \mathrm{ps}, F_{1{,}030} = 0, F_{800} = 0)$ without optical excitation. Concerning Figs. 2b and 3d, we define the relative switching efficiency as

$$E_s(\Delta t_{\mathrm{p-p}}) = 1 - \frac{(I_{8\times 2}(\Delta t_{\mathrm{p-p}}) - \langle I_{8\times 2}(\Delta t_{\mathrm{p-p}} > \Delta t^{*}_{\mathrm{p-p}})\rangle)}{\langle I_{8\times 2}(\Delta t_{\mathrm{p-p}} > \Delta t^{*}_{\mathrm{p-p}})\rangle} \qquad (1)$$

with $\Delta t^{*}_{\mathrm{p-p}} = 10$ ps (17 ps) in Fig. 2b (3d), respectively. In all cases, $\Delta t^{*}_{\mathrm{p-p}}$ is significantly larger than the temporal overlap of the two optical pulses given by their cross-correlation, and the damping constant of the coherent phonon oscillations.

Fourier analysis

We use super-Gaussian windows in the time domain to isolate the relevant sections in our datasets and reduce numerical artefacts of the fast Fourier transform:

$$F_{\mathrm{filt},t} = \exp\left(-\left(\frac{(t-t_{\mathrm{shift}})^2}{2\sigma_t^2}\right)^3\right) \tag{2}$$

The values of σ_t and t_{shift} used to create the respective figures are as follows: Fig. 2b, $\sigma_t = 3.2$ ps, $t_{\mathrm{shift}} = 4.5$ ps; Fig. 3d, $\sigma_t = 4.9$ ps, $t_{\mathrm{shift}} = 6.5$ ps; Fig. 3e, $\sigma_t = 3.5$ ps, $t_{\mathrm{shift}} = 4$ ps. To extract the contributions of the individual modes to the signal from Fig. 2b, a super-Gaussian frequency window

$$F_{\mathrm{filt},f} = \exp\left(-\left(\frac{(f-f_{\mathrm{c}})^2}{2\sigma_f^2}\right)^3\right) \tag{3}$$

is used to filter the relevant frequency range in the fast Fourier transforms. The data shown in Fig. 2b (bottom) are obtained by an inverse Fourier transform of the filtered Fourier transform (shear mode: $f_{\mathrm{c}} = 0.5$ THz, $\sigma_f = 0.10$ THz (frequency range 0.37–0.63 THz); rotation mode: $f_{\mathrm{c}} = 0.9$ THz, $\sigma_f = 0.07$ THz (frequency range 0.80–0.99 THz); d.c.: $f_{\mathrm{c}} = 0.0$ THz, $\sigma_f = 0.14$ THz (frequency range 0–0.19 THz)). Here, f_{c} and σ_f denote the centre frequency and width of the respective Fourier window. To study the delay-dependent frequency change of both the shear and the rotation mode, we perform a short-time Fourier transform of the dataset depicted in Fig. 3d (bottom), again with a super-Gaussian window function in the time domain ($\sigma_t = 3.6$ ps (see equation (2)), yielding the data shown in Fig. 3d (top).

Reflectivity and absorption of the indium monolayer

To relate the reflectivity changes measured by OPP to the absorption of the atomic indium wires, we follow ref. [41] for the optical properties of an ultrathin layer on top of a dielectric substrate. In our case, the silicon substrate has an essentially real (and comparatively large) refractive index ($n_{\mathrm{s}} = 3.67 + 0.005i$; $\lambda = 800$ nm). For normal incidence, the reflection and transmission coefficients r_0 and t_0 of the bare substrate are the standard expressions:

$$r_0 = \frac{(1-n_{\mathrm{s}})}{(1+n_{\mathrm{s}})};\ t_0 = \frac{2}{(n_{\mathrm{s}}+1)}. \tag{4}$$

Thus, the reflected wave is phase-shifted by 180°, and the transmitted wave is not phase-shifted. Furthermore, since the sheet conductivity σ^{s} of a monolayer satisfies $|Z_0\sigma^{\mathrm{s}}| \ll |n-1|$ (where Z_0 is free-space impedance), the monolayer-induced changes in transmission and reflection are both proportional to the real part of σ^{s}, as is the absorption A of the layer:

$$\Delta R = |r_0|^2\left(\frac{4}{n_{\mathrm{s}}^2-1}\right)\mathrm{Re}(Z_0\sigma^{\mathrm{s}}) = -\frac{4(1-n_{\mathrm{s}})}{(1+n_{\mathrm{s}})^3}\mathrm{Re}(Z_0\sigma^{\mathrm{s}}), \tag{5}$$

$$\Delta T = -\mathrm{Re}(n_{\mathrm{s}})|t_0|^2\left(\frac{2}{1+n_{\mathrm{s}}}\right)\mathrm{Re}(Z_0\sigma^{\mathrm{s}}) = -\mathrm{Re}(n_{\mathrm{s}})\frac{8}{(1+n_{\mathrm{s}})^3}\mathrm{Re}(Z_0\sigma^{\mathrm{s}}), \tag{6}$$

$$A = \frac{4}{|1+n_{\mathrm{s}}|^2}\mathrm{Re}(Z_0\sigma^{\mathrm{s}}). \tag{7}$$

In other words, owing to the large real and very small imaginary part of the substrate refractive index, the imaginary part of the sheet conductivity leads to only negligible (quadrature) components in the reflected and transmitted waves. The presence of the monolayer results in a ratio of reflectance change to layer absorption of $\Delta R/A = (n_{\mathrm{s}}-1)/(n_{\mathrm{s}}+1) = 0.57$. For the mechanism of absorption modulation (active for the rotation mode), pump-induced variations of the sheet conductivity $\delta\sigma^{\mathrm{s}}$ by coherent phonons will induce variations in reflectance (δR) and absorption (δA) following the same ratio $\delta R/\delta A = \Delta R/A$. Thus, transient reflectivity (OPP) directly measures the impact of a specific phonon mode on absorption.

Relating ULEED and OPP data

From the above, variations of layer absorption lead to proportional changes in reflectance, with a prefactor that depends on the total absorption of the monolayer. This allows us to estimate the monolayer absorption, assuming absorption modulation as the sole mechanism for the rotation mode. At an identical fluence of the first excitation (see Fig. 3d, e), for the rotation mode oscillation, we measure relative changes in reflectance $(\delta R/R)_{\mathrm{rot}} = 8\times10^{-5}$ and modulations of the ULEED intensity δI of 0.8% (intensity I normalized to value at negative times). The steepness of the fluence-dependent intensity (see Fig. 1c) $F_{\mathrm{th}} \times (\mathrm{d}I/\mathrm{d}F)_{F,\mathrm{th}} \approx 1.7$ at the threshold fluence $F_{\mathrm{th}} = 1$ mJ cm^{-2} is used to determine the relative changes in absorption of $\delta A/A = 0.47\%$ via $(\delta A/A)_{\mathrm{rot}} = (1/F_{\mathrm{th}}) \times ((\mathrm{d}I/\mathrm{d}F)|_{F,\mathrm{th}})^{-1} \times (\delta I/I)_{\mathrm{rot}}$. From these values, we obtain an estimate of the total absorption of the monolayer of

$$A \approx R \times \left(\frac{\delta R}{R}\right)_{\mathrm{rot}} \times \left(\frac{\delta A}{A}\right)_{\mathrm{rot}}^{-1} \times \left(\frac{\Delta R}{A}\right)^{-1} \approx 1\% \tag{8}$$

This value is of the same order as a recent estimate[14] (0.5%), again indicating that absorption modulation is a reasonable explanation for the rotation-mode contribution to the switching efficiency.

In turn, the observed differences in rotation and shear mode amplitudes between ULEED and OPP are sufficiently pronounced to imply a microscopic origin. In addition to possible ballistic contributions, this includes atomic-scale sample inhomogeneities such as local variations in barrier height.

Data availability

The data that support the findings of this study are available on request from the corresponding author.

47. Storeck, G., Vogelgesang, S., Sivis, M., Schäfer, S. & Ropers, C. Nanotip-based photoelectron microgun for ultrafast LEED. *Struct. Dyn.* **4**, 044024 (2017).
48. Van Hove, M. A., Weinberg, W. H. & Chan, C.-M. *Low-Energy Electron Diffraction: Experiment, Theory and Surface Structure Determination* (Springer, 1986).
49. Terada, Y. et al. Optical doping: active control of metal–insulator transition in nanowire. *Nano Lett.* **8**, 3577–3581 (2008).
50. Snijders, P. C. & Weitering, H. H. Colloquium: Electronic instabilities in self-assembled atom wires. *Rev. Mod. Phys.* **82**, 307–329 (2010).
51. Klasing, F. et al. Hysteresis proves that the In/Si(111) (8×2) to (4×1) phase transition is first-order. *Phys. Rev. B Condens. Matter Mater. Phys.* **89**, 121107 (2014).

Acknowledgements This work was funded by the European Research Council (ERC Starting Grant 'ULEED', ID: 639119) and the Deutsche Forschungsgemeinschaft (SFB-1073, project A05). We acknowledge discussions with N. S. Kozák, H. Schwoerer, R. Ernstorfer, M. Wolf, A. M. Wodtke and M. Horn-von Hoegen.

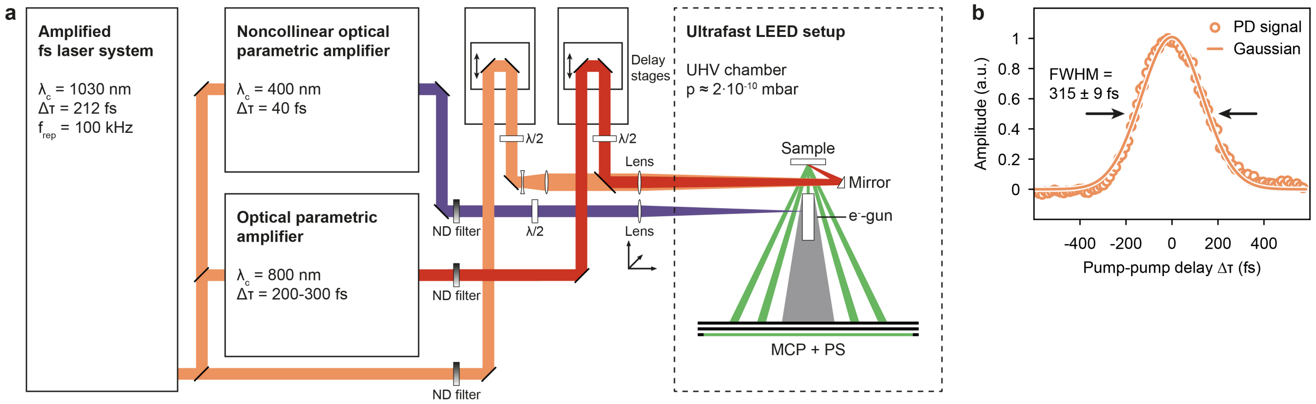
Author contributions The project was conceived by C.R., with contributions from J.G.H. Experiments and data analysis were conducted by J.G.H., with contributions from H.B., B.W., G.S. and F.K. The manuscript was written by J.G.H., H.B. and C.R. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

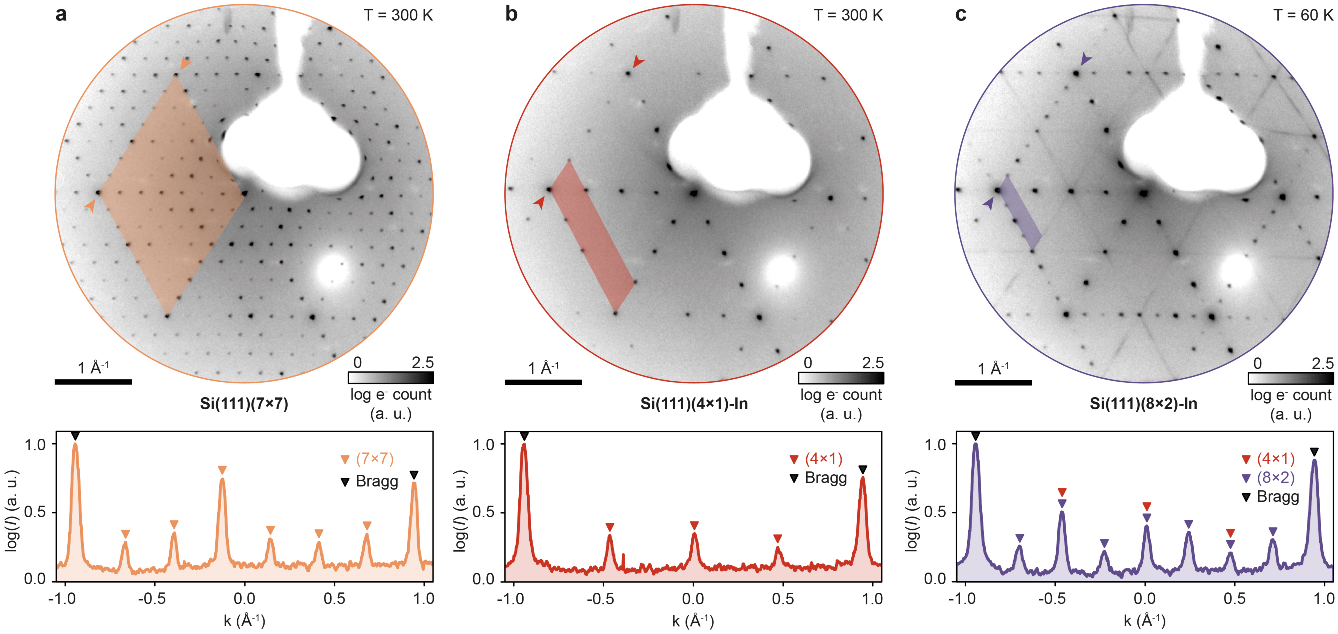
Additional information

Correspondence and requests for materials should be addressed to C.R.

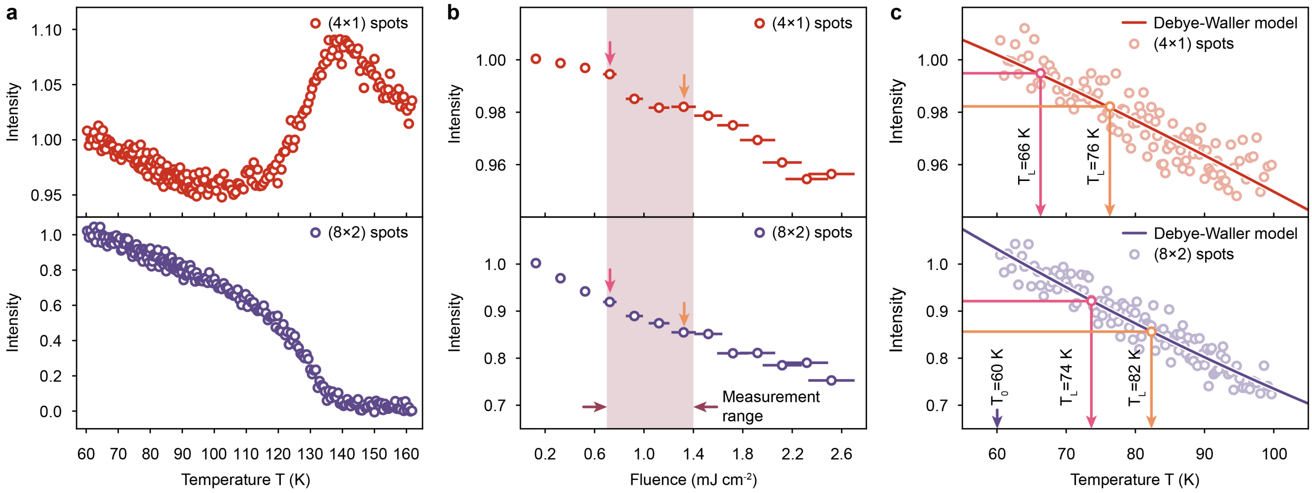
Peer review information *Nature* thanks Stefan Wippermann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



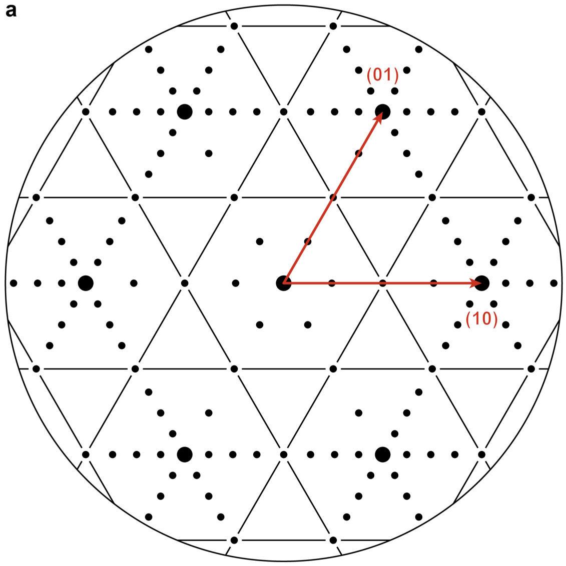
Extended Data Fig. 1 | ULEED set-up. a, Ultrashort laser pulses (P_1: $\lambda_c = 1{,}030$ nm, $\Delta\tau = 212$ fs) from an Yb:YAG amplifier (left) pump a non-collinear OPA (output: $\lambda_c = 400$ nm, $\Delta\tau = 40$ fs) and an OPA (output: P_2, $\lambda_c = 800$ nm, $\Delta\tau = 232$ fs). The 1,030-nm and 800-nm beams are independently attenuated and collinearly focused onto the sample by a single lens (400 mm focal length). The relative on-axis position of the two foci is controlled by adjusting the divergence of the 1,030-nm beam. The ultraviolet pulses are focused onto the tungsten needle emitter inside the electron gun (e^--gun) to generate ultrashort electron pulses. The relative timing between the electron probe and each of the two optical pump pulses is controlled independently by two separate optical delay stages. The pump-induced changes in the LEED pattern are recorded using a microchannel plate assembly. **b**, Cross-correlation of the two pump pulses recorded with a nonlinear photodiode to determine the temporal resolution of the double-pump experiment.



Extended Data Fig. 2 | Diffraction images. a–**c**, Diffraction images and lineouts of the clean (7 × 7)-reconstructed Si(111) surface (**a**), the (4 × 1) phase (**b**) and the (8 × 2) phase (**c**) recorded in our ultrafast LEED set-up ($E_{\mathrm{kin}} = 130$ eV). Coloured areas correspond to the unit cells in reciprocal space, arrows indicate the location of the lineouts shown below. In the transformation from the (4 × 1) to the (8 × 2) phase, the unit cell is doubled in both dimensions. The twofold streaks in the diffraction pattern of the (8 × 2) phase originate from a weak coupling between the atomic chains. The diffraction patterns of the indium-reconstructed phases feature contributions from three domains rotated by 120° with respect to each other, as the hexagonal structure of the underlying substrate allows for three different orientations of the atomic indium chains.



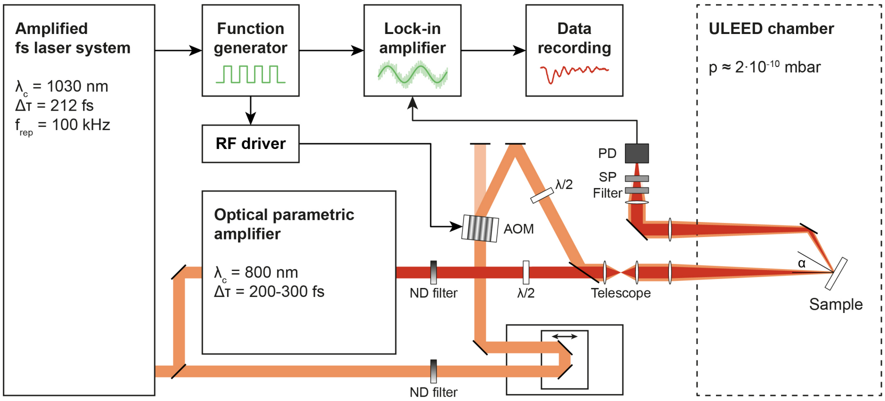
Extended Data Fig. 3 | Temperature calibration. **a**, Temperature-dependent integrated intensities of (4 × 1) (top) and (8 × 2) (bottom) diffraction spots across the phase transition ($T_c \approx 125$ K). **b**, Integrated diffraction spot intensities for $\Delta t_{\text{p-el}} < 0$ in Fig. 1c as a function of incident fluence. **c**, Temperature calibration: a Debye–Waller model is fitted to the diffraction spot intensities in **a** for temperatures in the range 60 K < T < 100 K. Comparing the suppressions in **b** and **c**, we find a maximum temperature increase $\Delta T_b \approx 22$ K for the highest fluence value ($F_{\max} \approx 1.35$ mJ cm^2) within our measurement range. Note that the resulting base temperature $T_b = 82$ K is well below the T_c.



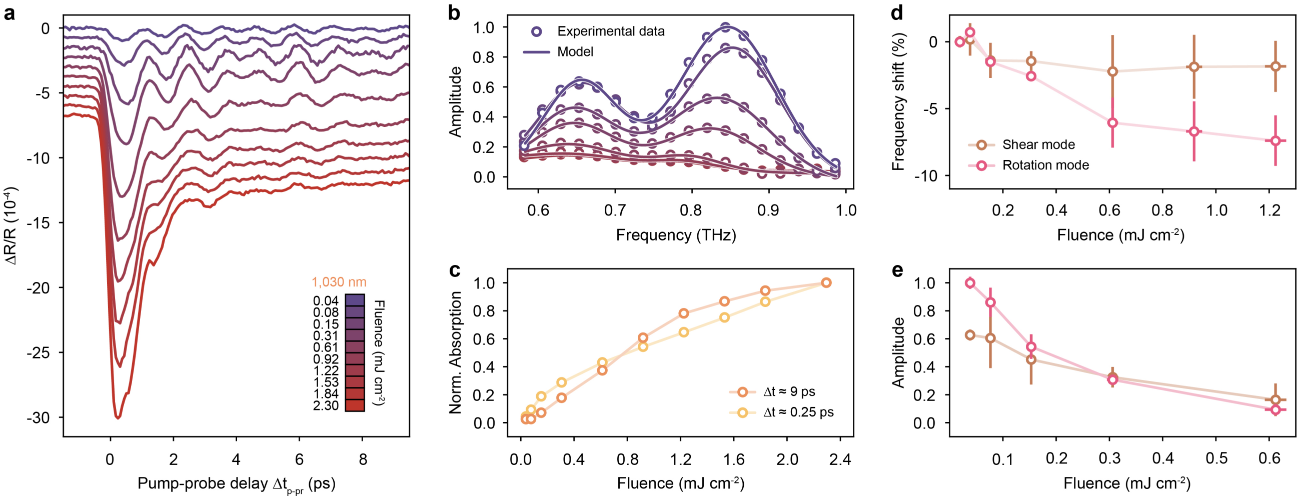
b

Pump-probe measurements (Fig.1) and static heating (Extended Data Fig.3)		
(4×1)	[(1 3), (2 2), (4 -4), (2 -4), (-2 -2), (0 -4), (0 -2), (0 -1), (-2 2)]	×1/4
(8×2)	[(3 5), (5 3), (8 -3), (5 -8), (3 -8), (-5 -3), (-8 -5)]	×1/8
Double-pulse measurements (Fig.2a)		
(4×1)	[(1 3), (2 2), (4 -4), (2 -4), (-2 -2), (3 1), (0 4)]	×1/4
(8×2)	[(3 5), (5 3), (8 -3), (5 -8), (3 -8), (-5 -3), (-8 -5), (-5 8)]	×1/8
Double-pulse measurements (Fig.2b and Fig.3d)		
(4×1)	[(1 3), (3 1), (4 0), (4 -1), (4 -4), (2 -4), (0 -4), (-1 -3), (-2 -2), (-3 -1), (-4 2), (-3 4), (-2 4), (-2 2), (-3 0), (-2 0), (0 -2), (0 -1), (2 -2), (3 -3)]	×1/4
(8×2)	[(3 5), (5 3), (8 -3), (8 -5), (3 -8), (5 -8), (3 -5), (-5 -3), (-8 3), (-8 5), (-5 8), (-3 8)]	×1/8

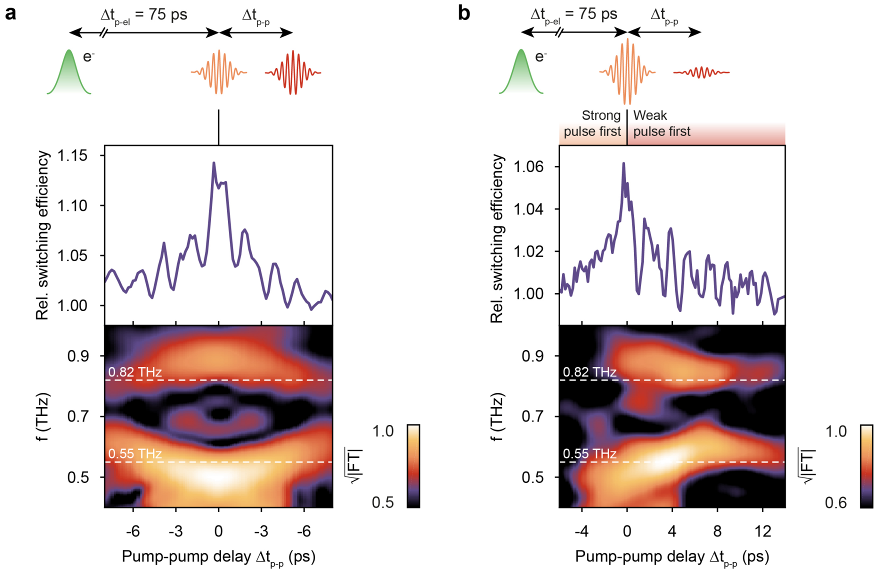
Extended Data Fig. 4 | Definition of basis vectors and diffraction spot indexing. a, Schematic LEED pattern of the (8 × 2) phase and basis vectors (red) of the reciprocal lattice used to index the diffraction spots. **b**, Complete list of diffraction spots used in analysis.



Extended Data Fig. 5 | Optical pump–probe set-up. a, Ultrashort laser pulses (P_1: $\lambda_c = 1{,}030$ nm, $\Delta\tau = 212$ fs, 'Pump') from an Yb:YAG amplifier (left) pump an OPA (output: P_2, $\lambda_c = 800$ nm, $\Delta\tau = 232$ fs, 'Probe'). The intensity of the pump beam is modulated at a frequency of 25 kHz by an acousto-optic modulator (AOM). Pump and probe beams are independently attenuated and collinearly focused onto the sample by a single lens (200-mm focal length). The relative on-axis position of the two foci can be adjusted using a telescope assembly. The reflected beams pass two short-pass filters (SP) blocking the pump pulses and are focused on a silicon photodiode (PD). The relative timing between pump and probe pulses is controlled by an optical delay stage. The pump-induced reflectivity changes of the sample are measured by processing the PD and reference signals in a lock-in amplifier. RF, radio-frequency; ND, neutral density.



Extended Data Fig. 6 | Ultrafast absorption modulation. a, Reflectivity change $\Delta R/R$ of the In/Si(111) surface as a function of the time-delay $\Delta t_{\text{p-pr}}$ between pump (1,030 nm) and probe pulses (800 nm; $F = 0.14$ mJ cm^{-2}). Offsets are added to the datasets for clarity. **b**, Fourier spectra of $\Delta R/R(\Delta t_{\text{p-pr}})$ for $F = 0.04$–1.22 mJ cm^{-2}, revealing two main coherent contributions ($f_1 = 0.65$ THz, $f_2 = 0.84$ THz for $F = 0.04$ mJ cm^{-2}) to the signals in **a**, attributed to the symmetric shear and rotation modes. An additional but minor lower-frequency contribution to the reflectivity cannot be excluded at this point, given the frequency resolution of the experiment. **c**, Transient ($\Delta t_{\text{p-pr}} \approx 0.25$ ps) and long-lived ($\Delta t_{\text{p-pr}} \approx 9$ ps) contributions to $\Delta R/R$ as a function of pump fluence. The data are normalized to $\Delta R/R(\Delta t_{\text{p-pr}} < 0)$ and the respective values for $F = 2.30$ mJ cm^{-2}. **d**, Fluence-dependent frequency shifts of the two modes. The rotation mode softens significantly for higher fluences (error bars, 95% CI of the fit). **e**, Normalized Fourier amplitudes of shear and rotation modes as a function of fluence.



Extended Data Fig. 7 | Short-time Fourier transforms. a, Relative switching efficiency as a function of the double-pulse delay $\Delta t_{\mathrm{p-p}}$ (top) and short-time Fourier transform (bottom) for equal pump pulses ($F_{1,030} = 0.32$ mJ cm^{-2}; $F_{800} = 0.21$ mJ cm^{-2}), revealing a pronounced softening/hardening of the shear/rotation component close to $\Delta t_{\mathrm{p-p}} = 0$ (see also Fig. 2b). **b**, Relative switching efficiency and short-time Fourier transform for unequal pump pulses ($F_{1,030} = 0.48$ mJ cm^{-2}; $F_{800} = 0.15$ mJ cm^{-2}, see also Fig. 3d).

A mobile robotic chemist

https://doi.org/10.1038/s41586-020-2442-2

Received: 1 November 2019

Accepted: 25 March 2020

Benjamin Burger[1], Phillip M. Maffettone[1], Vladimir V. Gusev[1], Catherine M. Aitchison[1], Yang Bai[1], Xiaoyan Wang[1], Xiaobo Li[1], Ben M. Alston[1], Buyi Li[1], Rob Clowes[1], Nicola Rankin[1], Brandon Harris[1], Reiner Sebastian Sprick[1] & Andrew I. Cooper[1]✉

Technologies such as batteries, biomaterials and heterogeneous catalysts have functions that are defined by mixtures of molecular and mesoscale components. As yet, this multi-length-scale complexity cannot be fully captured by atomistic simulations, and the design of such materials from first principles is still rare[1–5]. Likewise, experimental complexity scales exponentially with the number of variables, restricting most searches to narrow areas of materials space. Robots can assist in experimental searches[6–14] but their widespread adoption in materials research is challenging because of the diversity of sample types, operations, instruments and measurements required. Here we use a mobile robot to search for improved photocatalysts for hydrogen production from water[15]. The robot operated autonomously over eight days, performing 688 experiments within a ten-variable experimental space, driven by a batched Bayesian search algorithm[16–18]. This autonomous search identified photocatalyst mixtures that were six times more active than the initial formulations, selecting beneficial components and deselecting negative ones. Our strategy uses a dexterous[19,20] free-roaming robot[21–24], automating the researcher rather than the instruments. This modular approach could be deployed in conventional laboratories for a range of research problems beyond photocatalysis.

The mobile robot platform is shown in Fig. 1a and Extended Data Fig. 1. It can move freely in the laboratory and locates its position using a combination of laser scanning coupled with touch feedback for fine positioning (Methods and Supplementary Video 1). This gave an (x, y) positioning precision of ±0.12 mm and an orientation precision of $\theta \pm 0.005°$ within a standard laboratory environment with dimensions 7.3 m × 11 m (Fig. 1b; Extended Data Fig. 2; Supplementary Figs. 1–10). This precision allows the robot to carry out dexterous manipulations at the various stations in the laboratory (Fig. 1; Extended Data Fig. 3) that are comparable to those performed by human researchers, such as handling sample vials and operating instruments. The robot has human-like dimensions and reach (Fig. 1a, d) and it can therefore operate in a conventional, unmodified laboratory. Unlike many automated systems that can dispense only liquids, this robot dispenses both insoluble solids and liquid solutions with high accuracy and repeatability (Supplementary Figs. 12, 13, 16–20), broadening its utility in materials research. Factoring in the time needed to recharge the battery, this robot can operate for up to 21.6 h per day with optimal scheduling. The robot uses laser scanning and touch feedback, rather than a vision system. It can therefore operate in complete darkness, if needed (Supplementary Video 2), which is an advantage when carrying out light-sensitive photochemical reactions, as here. The robot arm and the mobile base comply with safety standards for collaborative robots, allowing human researchers to work within the same physical space (Supplementary Information section 1.5). A video of the robot operating an autonomous experiment over a 48-h period is shown in Supplementary Video 1.

The benefits of combining automated experimentation with a layer of artificial intelligence (AI) have been demonstrated for flow reactors[25], photovoltaic films[13], organic synthesis[8–10,14], perovskites[26] and in formulation problems[18]. However, so far no approaches have integrated mobile robotics with AI for chemical experiments. Here, we built Bayesian optimization[16–18] into a mobile robotic workflow to conduct photocatalysis experiments within a ten-dimensional space. Semiconductor photocatalysts that promote overall water splitting to produce both hydrogen and oxygen are still quite rare[15]. For many catalysts, a sacrificial hole scavenger is needed to produce hydrogen from water, such as triethylamine (TEA)[27] or triethanolamine (TEOA)[28], and these amines are irreversibly decomposed in the reaction. It has proved difficult to find alternative hole scavengers that compete with these organic amines[29].

Our objective was to identify bioderived hole scavengers with efficiencies that match petrochemical amines and that are not irreversibly decomposed, with the long-term aim of developing reversible redox shuttles. The photocatalyst that we chose was P10, a conjugated polymer that shows good HERs in the presence of TEOA[28]. We first used the robot to screen 30 candidate hole scavengers (Extended Data Fig. 4). This was done using a screening approach, without any AI. Initially, the robot loads a solid-dispensing station that weighs any solid components into sample vials (Fig. 1c), in this case the catalyst, P10. Next, the vials are transported 16 at a time in a rack to a dual liquid-dispensing station (Extended Data Fig. 3c), where the liquid components are added; here, 50 g l^{-1} aqueous solutions of the candidate hole scavengers (Supplementary Videos 3, 4). The robot then places the vials into a capping station, which caps the vials under nitrogen (Supplementary Fig. 21; Supplementary Video 5). Optionally, the capped vials are then placed into a sonication station (Supplementary Fig. 23; Supplementary

---

[1]Leverhulme Centre for Functional Materials Design, Materials Innovation Factory and Department of Chemistry, University of Liverpool, Liverpool, UK. ✉e-mail: aicooper@liverpool.ac.uk

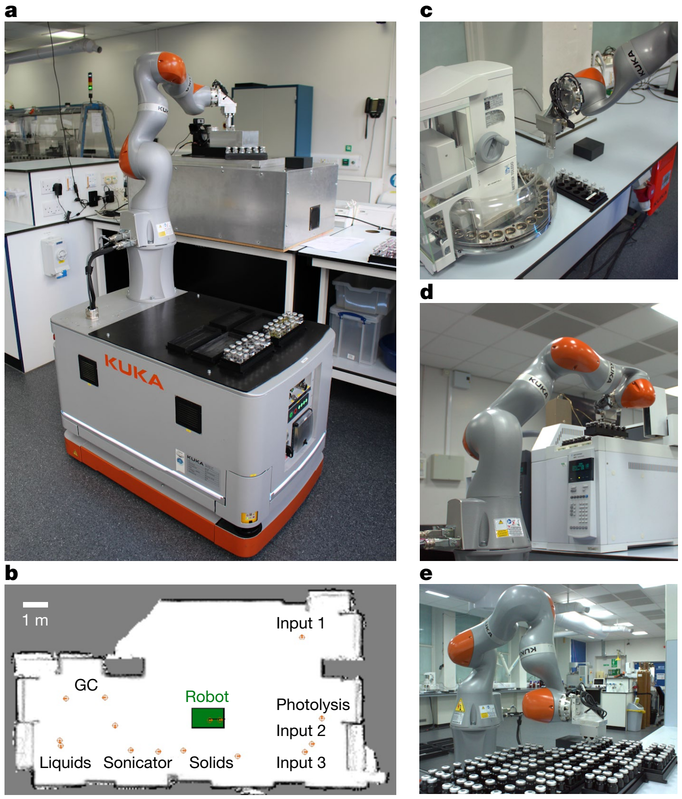


Fig. 1 | Autonomous mobile robot and experimental stations. **a**, Photograph showing robot loading samples into the photolysis station. **b**, Map of the laboratory generated by laser scanning showing positions of the eight stations; the orange crosshairs indicate recorded navigation locations and the robot position is indicated by the green rectangle. Inputs 1–3 are areas for the storage of empty vials or completed sample racks. GC, gas chromatography station. **c**, Robot loading empty sample vials into the solid-dispensing station before dispensing the photocatalyst. **d**, Loading the gas chromatography station with a new rack of samples for analysis. **e**, Storing racks of completed samples in Input Station 1 after gas chromatography analysis.

Video 3) to disperse the solid catalyst in the aqueous phase. The vials are then transported to a photolysis station, where they are illuminated with a mixture of ultraviolet and visible light (Fig. 1a; Extended Data Fig. 3b; Supplementary Fig. 24; Supplementary Video 6). After photolysis, the robot transfers the vials to a head space gas chromatography station where the gas phase is analysed for hydrogen (Fig. 1d) before storage of completed samples (Fig. 1e). Except for the capping station and the photolysis station, which were built specifically for this workflow, the other stations used commercial instruments with no physical hardware modifications: the robot operates them in essentially the same way that a human researcher would.

Conditional automation was used in this hole scavenger screen to repeat any hits; that is, samples that showed a hydrogen evolution rate (HER) of >200 μmol g^{-1} h^{-1} were automatically re-analysed five times. Most of the 30 scavengers produced little or no hydrogen (Extended Data Fig. 4), except for L-ascorbic acid (256 ± 24 μmol g^{-1} h^{-1}) and L-cysteine (1,201 ± 88 μmol g^{-1} h^{-1}). Analysis by ^{1}H nuclear magnetic resonance (NMR) spectroscopy showed that L-cysteine was cleanly converted to L-cystine (Supplementary Fig. 32), indicating that it may have potential as a reversible redox shuttle in an overall water splitting scheme[30].

While it showed promise as a hole scavenger, L-cysteine produced much less hydrogen than an aqueous solution of TEOA at the same gravimetric concentration (2,985 ± 103 μmol g^{-1} h^{-1} at 50 g l^{-1}). We therefore sought to increase the HER of the P10/L-cysteine system by using an autonomous robotic search based on five hypotheses (Fig. 2a).

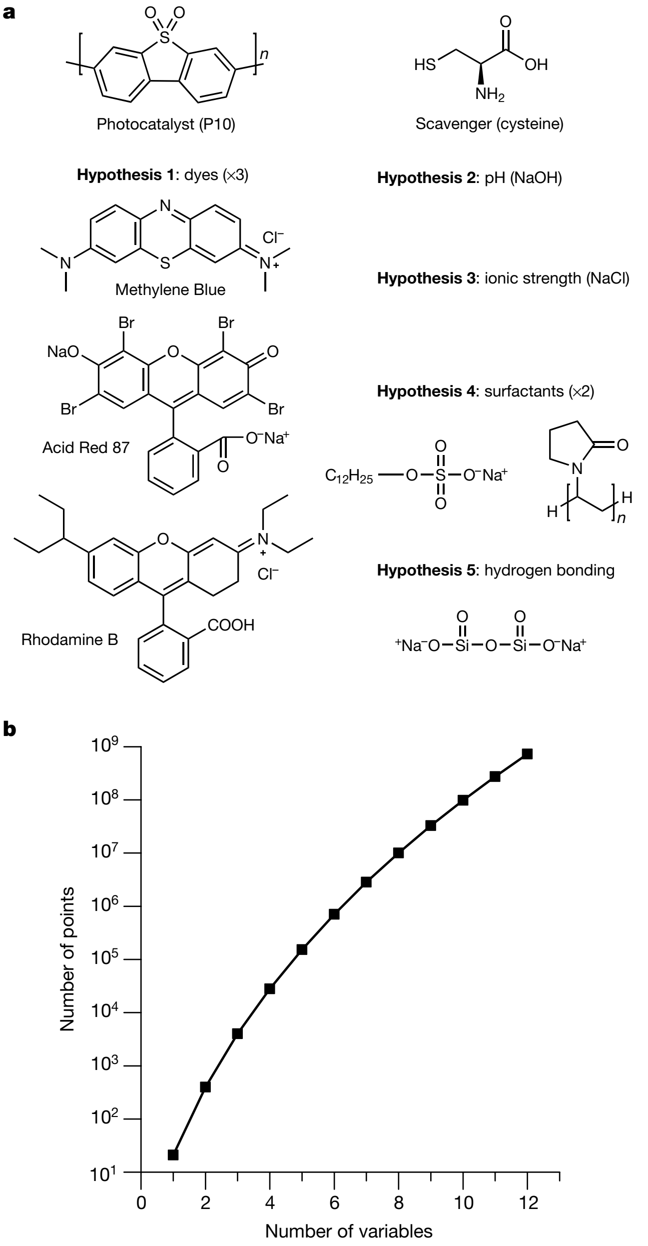


Fig. 2 | Hypothesis-led autonomous search strategy. **a**, The robot searches chemical space to optimize the activity of the photocatalyst + scavenger combination according to five separate hypotheses. It does this by simultaneously varying the concentration of the ten chemical species shown here. **b**, Plot showing the size of the simplex, or the search space, created with a discretization of 19 concentrations for each liquid and 21 concentration levels for the solid catalyst, P10, which corresponds to the solid/liquid dispensing precision over the constrained space of the experiment. For this ten-component problem, the full simplex has 98,423,325 points.

The first hypothesis was that dye sensitization might improve light absorption and hence the HER, as found for the structurally related covalent organic framework, FS-COF[31]. Here, three dyes were investigated (Rhodamine B, Acid Red 87[31] and Methylene Blue). Second, we hypothesized that pH might influence the catalytic activity (NaOH addition). The third hypothesis was that ionic strength could also be important[32] (NaCl addition). Catalyst wettability is known to be a factor in photocatalytic hydrogen evolution using conjugated polymers[33], so the addition of surfactants (sodium dodecyl sulphate, SDS, and

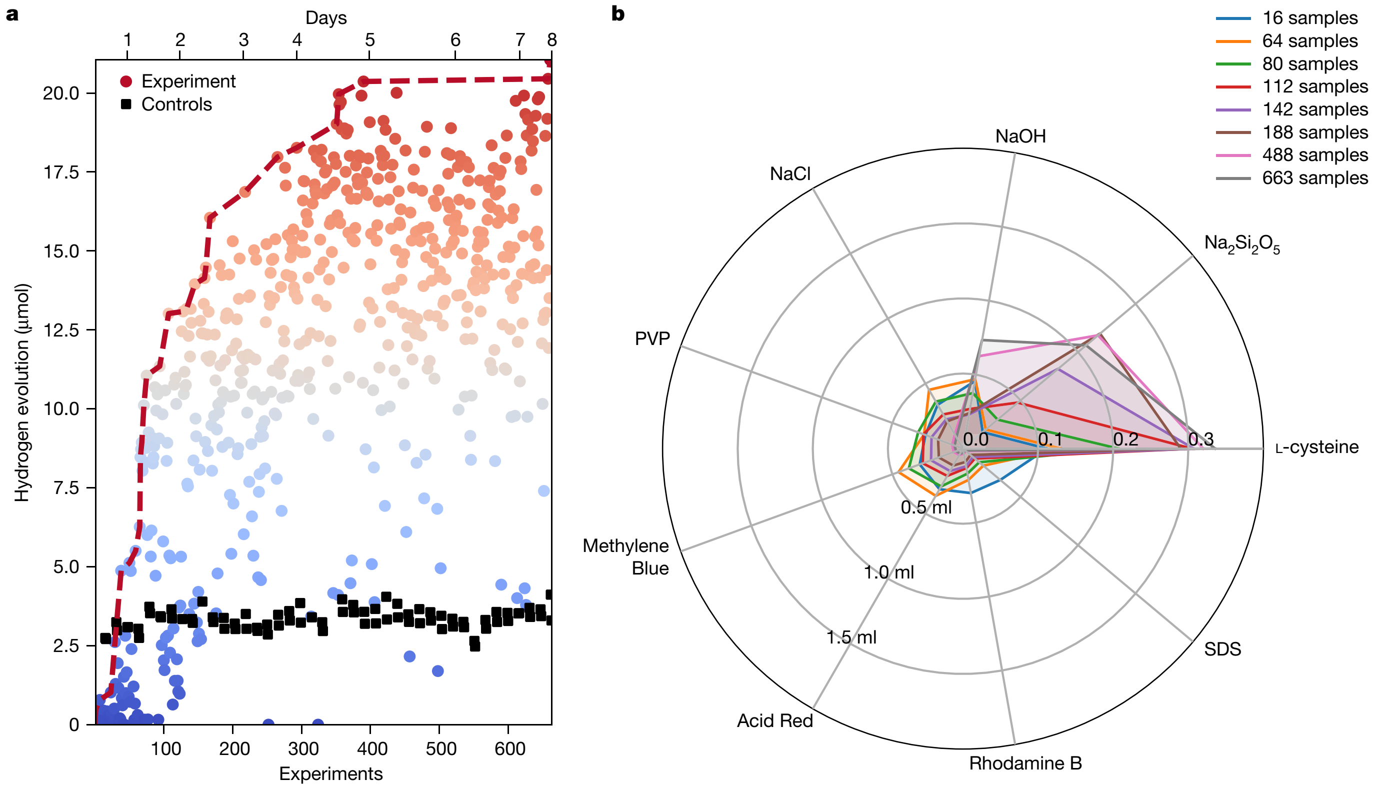


Fig. 3 | Output from the autonomous robotic search. a, Plot showing hydrogen evolution achieved per experiment in an autonomous search that extended over 8 days. Sixteen experiments were performed per batch, along with two baseline controls. The baseline hydrogen evolution was 3.36 ± 0.30 µmol (black squares). The maximum rate attained after 688 experiments was 21.05 µmol h^{-1}. The robot made 319 moves between stations and travelled a total distance of 2.17 km during this 8-day experiment. **b**, Radar plot showing the evolution of the average sampling of the search space in millilitres; the scale denotes the fraction of maximum solution volume dispensed. The starting conditions (Batch 1) were chosen randomly. The best catalyst formulation found after 43 batches contained P10 (5 mg), NaOH (6 mg), L-cysteine (200 mg) and $Na_2Si_2O_5$ (7.5 mg) in water (5 ml).

polyvinylpyrrolidone, PVP) formed our fourth hypothesis. Fifth, we speculated that sodium disilicate might act as a hydrogen-bonding anchor for the scavenger, L-cysteine, or for the dyes, based on the observation that it aids in the absorption of dyes onto the surface of carbon nitride[34].

These five hypotheses had the potential to be synergistic or anti-synergistic; for example, ionic strength could either enhance or decrease dye absorption onto the surface of the photocatalyst. We therefore chose to explore all five hypotheses at once. This involved the simultaneous variation of the concentration of P10, L-cysteine, the three dyes, NaOH, NaCl, the two surfactants, and sodium disilicate, which equates to a ten-variable search space (Fig. 2a). The space was constrained by the need to keep a constant liquid volume (5 ml) and therefore head space for gas chromatography analysis and by the minimal resolution for liquid dispensing module (0.25 ml) and solid dispensing module (0.2 mg).

Problems of this type are defined by a simplex that scales exponentially with size (Fig. 2b). For this specific search space, there were more than 98 million points. Full exploration of such a space is unfeasible, so we developed an algorithm that performs Bayesian optimization based on Gaussian process regression and parallel search strategy[35] (see Methods). To generate a new batch, we build a surrogate model predicting the HER of potential formulations based on the measurements performed so far and quantify the uncertainty of prediction. Subsequent sampling points are chosen using a capitalist acquisition strategy, where a portfolio of upper confidence bound functions is generated on an exponential distribution of greed to create markets of varying risk aversion, which are searched for global maxima. Each market is given an agent that searches to return a global maximum, or batch of *k*-best maxima. The uneven distribution of greed allows some suggested points to be highly exploitative, some to be highly explorative, and most to be balanced, thus making the strongest use of the parallel batch experiments.

The output from this autonomous robotic search is shown in Fig. 3a. The baseline HER for P10 and L-cysteine only (5 mg P10 in 5 ml of 20 g l^{-1} L-cysteine) was 3.36 ± 0.30 µmol h^{-1}. Given that the robot would operate autonomously over multiple days, this two-component mixture was repeated throughout the search (two samples per batch) to check for long-term experimental stability (black squares in Fig. 3a). Initially, the robot started with random conditions and discovered multicomponent catalyst formulations that were mostly less active than P10 and L-cysteine alone (the first 22 experiments in Fig. 3a). The robot then discovered that adding NaCl provides a small improvement to the HER, validating the hypothesis that ionic strength is important. In the same period, the robot found that maximizing both P10 and L-cysteine increased the HER. In further experiments (15–100), the robot discovered that none of the three dyes or the two surfactants improves the HER; indeed, they are all detrimental, counter to our first and fourth hypotheses. These five components were therefore deselected after around 150 experiments (Fig. 4); that is, after about 2 days in real experimental time (Fig. 3a). Here, P10 differs from the structurally related crystalline fused-sulfone covalent organic framework (FS-COF), where the addition of Acid Red 87 increased the HER[31]. After 30 experiments, the robot learned that adding sodium disilicate improves the HER substantially in the absence of dyes (up to 15 µmol after 300 experiments), while deprioritizing the addition of NaOH and NaCl. After 688 experiments, which amounted to 8 days of autonomous searching, the robot found that the optimum catalyst formulation is a

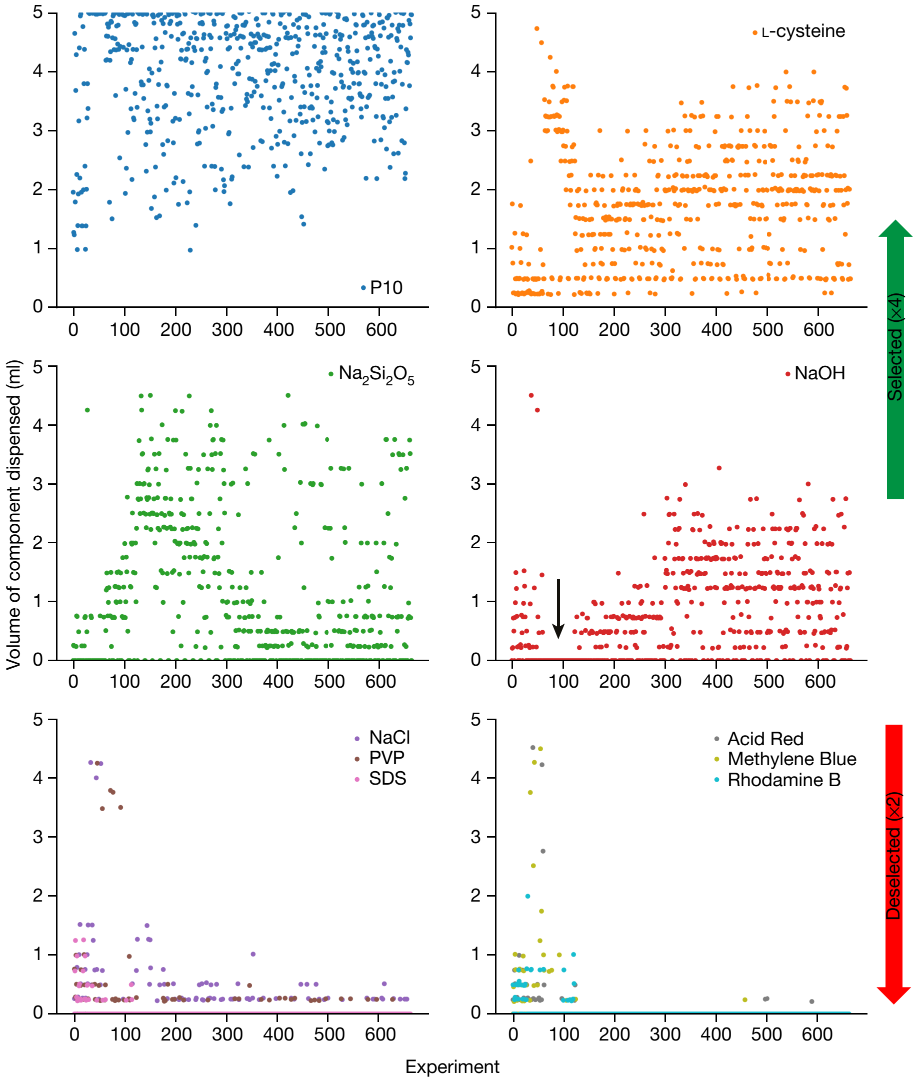


Fig. 4 | Selection and deselection of photocatalyst formulation components. Plots showing the mass (in milligrams for P10) or volume (millilitres for all other components) dispensed for the various components in the search space as a function of experiment. The photocatalyst, P10, and the scavenger, L-cysteine, are selected, along with sodium disilicate ($Na_2Si_2O_5$) and NaOH. All other components were deselected after around 150 experiments. The three dyes and the two surfactants had a negative effect on the HER. NaCl had a small positive effect, but less so than the four selected components, and it was therefore deselected. Note that NaOH was initially deselected, and not included in experiments 15–283 (see black arrow), while $Na_2Si_2O_5$ and L-cysteine were favoured. The positive effect of NaOH was initially masked by negative components such as the dyes. Later in the search, NaOH was favoured, ultimately in preference to $Na_2Si_2O_5$, illustrating the benefit of using an uneven distribution of greed in the search.

mixture of NaOH, L-cysteine, sodium disilicate and P10, giving a HER of 21.05 µmol h^{-1}, which was six times higher than the starting conditions.

A number of scientific conclusions can be drawn from these data. Increased ionic strength is beneficial for hydrogen production (NaCl addition), but not as beneficial as increasing the pH (NaOH/sodium disilicate addition), which also increases the ionic strength. We had not investigated surfactant addition before, but for the two surfactants studied here, at least, the effect on catalytic activity is purely negative. Intriguingly, the dye sensitization that we observed for a structurally similar covalent organic framework, FS-COF[31], does not translate to this polymer, P10, possibly because the COF is porous whereas P10 is not.

To explore the dependence of the algorithmic search performance on the random starting conditions, we carried out 100 in silico virtual searches, each with a different random starting point, using a regression model and random noise to return virtual results (Supplementary Information section 7). Around 160 virtual experiments were needed, on average, to find solutions with 95% of the global maximum HER (Fig. 5).

We estimate that it would have taken a human researcher several months to explore these five hypotheses in the same level of detail using standard, manual approaches (Supplementary Fig. 31). Manual hydrogen evolution measurements require about 0.5 days of researcher time per experiment (1,000 experiments take 500 days). The semi-automated robotic methods that we developed recently[33] can perform 100 experiments per day (a half-day to set up, plus a half-day for automatic dispensing and measurement; 1,000 experiments take 10 days, of which 5 days are dedicated researcher time). The autonomous robot that we present here also requires half a day to set it up initially, but it then runs unattended over multiple days (1,000 experiments take 0.5 days of researcher time). Hence, the autonomous workflow is 1,000 times faster than manual methods, and at least ten times faster than semi-automated but non-autonomous robotic workflows. It is unlikely that a human researcher would have persevered with this multivariate experiment using manual approaches given that it might have taken 50 experiments or 25 days to locate even a modest enhancement in the HER (Fig. 3a). The platform allows us to tackle search spaces of a size that would otherwise be impossible, which is an advantage for problems where our current level of understanding does not allow us to reduce the number of candidate components to a more manageable number. There were ten components in the example given here, but search spaces with up to at least 20 components should be tractable with some modifications to the algorithm.

It took an initial investment of time to build this workflow (approximately 2 years), but once operating with a low error rate (Supplementary Fig. 38), it can be used as a routine tool. The time required to implement this approach in another laboratory would be much shorter, since much of the 2-year development timescale involved core protocols and software that are transferable to other research problems. Also, this modular approach to laboratory automation uses instruments in a physically unmodified form, so that it will be straightforward to add further modules, such as for NMR or X-ray diffraction, now that the basic

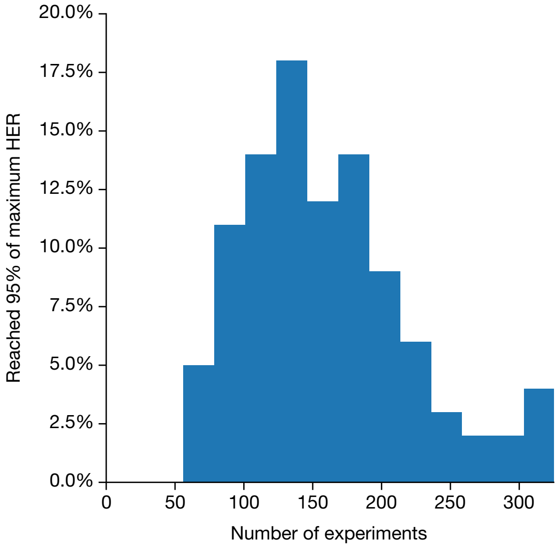


Fig. 5 | Virtual in silico experiments. Histogram showing the number of virtual experiments needed to reach 95% of the optimal HER, as determined by carrying out 100 in silico searches, each with a different random starting point.

principles are in place. This modularity makes our strategy applicable to a wide range of research problems beyond chemistry. The speed and efficiency of the method allow the exploration of large multivariate spaces, and the autonomous robot has no confirmation bias[36]; this raises the prospect of emergent function in complex, multi-component materials that we could not design in the conventional way. Autonomous mobile robots could also have extra advantages in experiments with especially hazardous materials, or where traceability and auditing are important, such as in pharmaceutical processes.

This approach also has some limitations. For example, the Bayesian optimization is blind, in that all components have equal initial importance. This robotic search does not capture existing chemical knowledge, nor include theory or physical models: there is no computational brain. Also, this autonomous system does not at present generate and test scientific hypotheses by itself[37]. In the future, we propose to fuse theory and physical models with autonomous searches: for example, computed structures and properties[1–5] could be used to bias searches towards components that have a higher likelihood of yielding the desired property. This will be important for search spaces with even larger numbers of components where purely combinatorial approaches may become inefficient. To give one example, energy–structure–function maps[38] could be computed for candidate crystalline components to provide Boltzmann energy weightings[39] for calculated properties, such as a charge transport or optical gap, to bias the robotic search.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2442-2.

1. Kang, K., Meng, Y. S., Bréger, J., Grey, C. P. & Ceder, G. Electrodes with high power and high capacity for rechargeable lithium batteries. *Science* **311**, 977–980 (2006).
2. Woodley, S. M. & Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **7**, 937–946 (2008).
3. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
4. Collins, C. et al. Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature* **546**, 280–284 (2017).
5. Davies, D. W. et al. Computer-aided design of metal chalcohalide semiconductors: from chemical composition to crystal structure. *Chem. Sci.* **9**, 1022–1030 (2018).
6. King, R. D. Rise of the robo scientists. *Sci. Am.* **304**, 72–77 (2011).
7. Li, J. et al. Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221–1226 (2015).
8. Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 15733 (2017).
9. Bédard, A.-C. et al. Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220–1225 (2018).
10. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
11. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
12. Langner, S. et al. Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multi-component systems. Preprint at https://arxiv.org/abs/1909.03511 (2019).
13. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. Preprint at https://arxiv.org/abs/1906.05398 (2019).
14. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
15. Wang, Z., Li, C. & Domen, K. Recent developments in heterogeneous photocatalysts for solar-driven overall water splitting. *Chem. Soc. Rev.* **48**, 2109–2125 (2019).
16. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & Freitas, N. D. Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* **104**, 148–175 (2016).
17. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
18. Roch, L. M. et al. ChemOS: orchestrating autonomous experimentation. *Sci. Robot.* **3**, eaat5559 (2018).
19. Chen, C.-L., Chen, T.-R., Chiu, S.-H. & Urban, P. L. Dual robotic arm "production line" mass spectrometry assay guided by multiple Arduino-type microcontrollers. *Sens. Actuat. B* **239**, 608–616 (2017).
20. Fleischer, H. et al. Analytical measurements and efficient process generation using a dual-arm robot equipped with electronic pipettes. *Energies* **11**, 2567 (2018).
21. Liu, H., Stoll, N., Junginger, S. & Thurow, K. Mobile robot for life science automation. *Int. J. Adv. Robot. Syst.* **10**, 288 (2013).
22. Liu, H., Stoll, N., Junginger, S. & Thurow, K. A fast approach to arm blind grasping and placing for mobile robot transportation in laboratories. *Int. J. Adv. Robot. Syst.* **11**, 43 (2014).
23. Abdulla, A. A., Liu, H., Stoll, N. & Thurow, K. A new robust method for mobile robot multifloor navigation in distributed life science laboratories. *J. Contrib. Sci. Eng.* **2016**, 3589395 (2016).
24. Dömel, A. et al. Toward fully autonomous mobile manipulation for industrial environments. *Int. J. Adv. Robot. Syst.* **14**, https://doi.org/10.1177/1729881417718588 (2017).
25. Schweidtmann, A. M. et al. Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. *Chem. Eng. J.* **352**, 277–282 (2018).
26. Zhi, L. et al. Robot-accelerated perovskite investigation and discovery (RAPID): 1. Inverse temperature crystallization. Preprint at https://doi.org/10.26434/chemrxiv.10013090.v1 (2019).
27. Matsuoka, S. et al. Photocatalysis of oligo (*p*-phenylenes): photoreductive production of hydrogen and ethanol in aqueous triethylamine. *J. Phys. Chem.* **95**, 5802–5808 (1991).
28. Shu, G., Li, Y., Wang, Z., Jiang, J.-X. & Wang, F. Poly(dibenzothiophene-*S*,*S*-dioxide) with visible light-induced hydrogen evolution rate up to 44.2 mmol h^{-1} g^{-1} promoted by K_2HPO_4. *Appl. Catal. B* **261**, 118230 (2020).
29. Pellegrin, Y. & Odobel, F. Sacrificial electron donor reagents for solar fuel production. *C. R. Chim.* **20**, 283–295 (2017).
30. Sakimoto, K. K., Zhang, S. J. & Yang, P. Cysteine–cystine photoregeneration for oxygenic photosynthesis of acetic acid from CO_2 by a tandem inorganic–biological hybrid system. *Nano Lett.* **16**, 5883–5887 (2016).
31. Wang, X. et al. Sulfone-containing covalent organic frameworks for photocatalytic hydrogen evolution from water. *Nat. Chem.* **10**, 1180–1189 (2018).
32. Schwarze, M. et al. Quantification of photocatalytic hydrogen evolution. *Phys. Chem. Chem. Phys.* **15**, 3466–3472 (2013).
33. Bai, Y. et al. Accelerated discovery of organic polymer photocatalysts for hydrogen evolution from water through the integration of experiment and theory. *J. Am. Chem. Soc.* **141**, 9063–9071 (2019).
34. Zhang, J. et al. H-bonding effect of oxyanions enhanced photocatalytic degradation of sulfonamides by g-C_3N_4 in aqueous solution. *J. Hazard. Mater.* **366**, 259–267 (2019).
35. Hutter, F., Hoos, H. H. & Leyton-Brown, K. *Parallel Algorithm Configuration* 55–70 (Springer, 2012).
36. Mynatt, C. R., Doherty, M. E. & Tweney, R. D. Confirmation bias in a simulated research environment: an experimental study of scientific inference. *Q. J. Exp. Psychol.* **29**, 85–95 (1977).
37. King, R. D. et al. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252 (2004).
38. Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
39. Campbell, J. E., Yang, J. & Day, G. M. Predicted energy–structure–function maps for the evaluation of small molecule organic semiconductors. *J. Mater. Chem. C* **5**, 7574–7584 (2017).

Methods

Robot specifications

The robot used was a KUKA Mobile Robot mounted on a KUKA Mobile Platform base (Fig. 1a; Extended Data Fig. 1). The robot arm has a maximum payload of 14 kg and a reach of 820 mm. The KUKA Mobile Platform base can carry payloads of up to 200 kg. The robot arm and the mobile base have a combined mass of approximately 430 kg. The movement velocity of the robot was restricted to 0.5 m s^{-1} for safety reasons (section 1.5 of the Supplementary Information,). A multipurpose gripper was designed to grasp 10-ml gas chromatograph sample vials, solid dispensing cartridges, and a 16-position sample rack (Extended Data Fig. 5), thus allowing a single robot to carry out all of the tasks required for this workflow. This robot was specified to be a flexible platform for a wide range of research tasks beyond those exemplified here; for example, the 14 kg payload capacity for the arm is not fully used in these experiments (one rack of filled vials has a mass of 580 g), but it could allow for manipulations such as opening and closing the doors of certain equipment. Likewise, the height and reach of the robot allows for operations such as direct loading of samples into the gas chromatograph instrument (Fig. 1d). By contrast, a smaller and perhaps less expensive robot platform might require an additional, dedicated robot arm to accomplish this, or inconvenient modifications to the laboratory, such as lowering bench heights.

Robot navigation

In a process analogous to simultaneous localization and mapping (SLAM)[40], the robot tracks a cloud of possible positions, and updates its position to the best fit between the output of its laser scanners and the map for each position in the cloud. The position of the robot is determined by x and y (its position on the map) and θ (its orientation angle). Histograms of the robot position measured over 563 movements are shown in Supplementary Figs. 2–5, which show that the (x, y) positioning precision was better than ±10 mm and the orientation precision was less than ±2.5°, as achieved within a real, working laboratory environment. This level of precision allows navigation to the various experimental stations in the laboratory, but it does not allow fine manipulations, such as placement of sample vials. The precision was therefore enhanced by using a touch-sensitive 6-point calibration method. Here, the robot touches six points on a cube that is associated with each experimental station to find the position and orientation of the cube relative to the robot (Supplementary Figs. 7–11). This increased the positioning precision to ±0.12 mm and the orientation precision to ±0.005°. This makes it possible for the robot to operate instruments and to carry out delicate manipulations such as vial placements at a level of precision that is broadly comparable to a human operator.

Experimental stations

The workflow comprised six steps, each with its own station. Solid dispensing was carried out with a Quantos QS30 instrument (Mettler Toledo) (Fig. 1c; Supplementary Fig. 11; Extended Data Fig. 3a; Supplementary Video 3). Liquid dispensing was carried out with a bespoke system that used a 200 series Mini Peristaltic Pump (Williamson) and a PCG 2500-1 scale (Kern), to dispense liquids gravimetrically using a feedback loop (section 2.2 in Supplementary Information; Supplementary Video 5). This system showed excellent precision and accuracy for a range of aqueous and non-aqueous liquids over 20,000 dispenses (Supplementary Figs. 16, 17, 19, 20). A bespoke instrument was built (Labman) to allow both for sample inertization (to exclude oxygen) and cap crimping in one step. It would be straightforward to modify this platform to allow other gases to be introduced; for example, to study photocatalytic CO_2 reduction. The instrument used caps from a vibratory bowl feeder to cap-crimp 10-ml headspace vials (section 2.2 in Supplementary Information; Supplementary Video 5). If required, a sonication station was used to disperse the solid photocatalyst in the aqueous solution, before reaction (Supplementary Fig. 23). Photolysis was carried out at a bespoke photolysis station (Fig. 1a) that uses vibration to agitate liquids and a light source that is composed of BL368 tubes and LED panels (Extended Data Fig. 5b; Supplementary Fig. 24; Supplementary Video 6). Gas chromatograph measurements were performed with a 7890B GC and a 7697A Headspace Sampler from Agilent GC (Supplementary Video 3; Extended Data Fig. 3d). The experimental stations were controlled by a process management system module, which contains all of the process logic for controlling the labware. Communication between the process management system and the stations was achieved using various communication protocols (TCP/IP over WIFI/LAN; RS-232), as detailed in section 2.7 in the Supplementary Information (Supplementary Fig. 28).

Autonomous search procedure and scheduling

The robot worked with batches of 16 samples per sample rack and ran 43 batches (688 experiments) during the search. Of these 688 experiments, 11 results were discarded because of workflow errors or because the system flagged that the oxygen level was too high (faulty vial seal). It took, on average, 183 min to prepare and photolyse each batch of samples and then 232 min per batch to complete the gas chromatograph analysis. The detailed timescales for each of the step in the workflow are shown in Extended Data Fig. 6. The work was heuristically scheduled in parallel, with the robot starting the oldest available scheduled job. While the robot was working on one job, other instruments, such as the solid dispenser, the photolysis station and the gas chromatograph, worked in parallel. This system can process up to six batches at once, but given the timescales for this specific workflow, where the preparation/reaction time is approximately equal to the analysis time, the robot processed two batches simultaneously. That is, it prepared samples and ran photolysis for one batch while analysing the hydrogen produced for the second batch using the gas chromatograph. The robot recharged its battery automatically in between two jobs when the battery charge reached a 25% threshold. The robot was charged but idle for approximately 32% of the time in this experiment, largely because of time spent waiting for the gas chromatograph analysis, which is the slow step. In principle, this time could be used to run other experiments in parallel. The autonomous workflow was programmed to alert the operator automatically when the system is out of stock (if, for example, it ran out of sample vials or stock solutions were low), or if a part of the workflow failed (section 8 of the Supplementary Information). Most errors could be reset remotely without being in the laboratory because all stations were equipped with 24/7 closed-circuit television cameras (Supplementary Fig. 39).

Bayesian search algorithm

The AI guidance for the autonomous mobile robot was a batched, constrained, discrete Bayesian optimization algorithm. Traditionally, Bayesian optimization is a serial algorithm tasked with finding the global maximum of an unknown objective function[16]. Here, this equates to finding the optimal set of concentrations in a multicomponent mixture for photocatalytic hydrogen generation. The algorithm builds a model that can be updated and queried for the most promising points to inform subsequent experiments. This surrogate model is constructed by first choosing a functional prior $\phi_{\mathrm{prior}}(\theta)$, informed by existing chemical knowledge (if any). Given data $\mathcal{D}$ and a likelihood model $\phi_{\mathrm{likelihood}}(\mathcal{D}|\theta)$, this yields a posterior distribution of models using Bayes' theorem:

$$\phi_{\mathrm{posterior}}(\theta|\mathcal{D}) = \frac{\phi_{\mathrm{likelihood}}(\mathcal{D}|\theta)\phi_{\mathrm{prior}}(\theta)}{\phi(\mathcal{D})} \tag{1}$$

The Gaussian process prior used a Matern similarity kernel, constant scaling and homoscedastic noise[41]. This composite kernel allows for variable smoothness, catalytic activity and experimental noise. The form and respective hyperparameters were refined using cross-validation on other, historical photocatalysis datasets (350 experiments). Other alternatives for a functional prior included Bayesian neural networks[17]; but Gaussian processes were selected here for robustness and flexibility[42]. An acquisition function, α_{UCB}, was assembled from the posterior distribution by considering the posterior mean, $\mu(x)$, and uncertainty, $\sigma(x)$. The maximum of this function was then used as the next suggested experiment. To balance exploitation (prioritizing areas where the mean is expected to be largest) and exploration (prioritizing areas where the model is most uncertain), we used an upper confidence bound that is dependent on a single hyperparameter, β, to govern how 'greedy' (exploitative) the search is:

$$\alpha_{\mathrm{UCB}}(x\,;\mathcal{D}) := \mu(x) + \beta\sigma(x) \tag{2}$$

The portfolio of acquisition functions for different values of β, which we call markets, was used to generate a batch. This 'capitalist' approach has the advantage of simple parallelization and is robust across variable batch sizes[35]. Our method allowed us to constrain the sum of all liquid components to 5 ml to allow a constant gas headspace volume for gas chromatograph analysis. The sum total volume constraint was handled during the market searches; discretization, which was determined by instrument resolution, was handled after the market searches. The market search was completed using a large initial random sampling followed by a batch of seeded local maximizations using a sequential least-squares programming (SLSQP) algorithm as implemented in the scipy.optimize package. This maximization occurs in a continuous space, and the results are placed into discrete bins following the experimental precision. The explored space is tracked as a continuous variable for model building and as a discrete variable for acquisition function maximization. The algorithm was implemented using the scikit-learn and in scipy packages[43].

Materials and synthetic procedures

The polymeric photocatalyst P10 was synthesized and purified according to a modification on a literature procedure[44] (section 10 of the Supplementary Information). For solid dispensing, the polymer was ground with mortar and pestle before use. Sodium disilicate was obtained as a free sample from Silmaco. Tap water was purified with PURELAB Ultra System. All other materials were purchased from Sigma-Aldrich and used as received.

Data availability

The implementation of the liquid-dispensing station, photolysis station and the workflow, along with three-dimensional designs for labware developed in the project, are available at https://bitbucket.org/ben_burger/kuka_workflow, the code for the robot at and the Bayesian optimizer is available at https://github.com/Taurnist/kuka_workflow_tantalus and https://github.com/CooperComputationalCaucus/kuka_optimizer. Additional design details can be obtained from the authors upon request.

40. Fuentes-Pacheco, J., Ruiz-Ascencio, J. & Rendón-Mancha, J. M. Visual simultaneous localization and mapping: a survey. *Artif. Intell. Rev*. **43**, 55–81 (2015).
41. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
42. Matthews, A. G. G., Rowland, M., Hron, J., Turner, R. E. & Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. Preprint at https://arxiv.org/abs/1804.11271 (2018).
43. Millman, K. J. & Aivazis, M. Python for scientists and engineers. *Comput. Sci. Eng*. **13**, 9–12 (2011).
44. Sachs, M. et al. Understanding structure-activity relationships in linear polymer photocatalysts for hydrogen evolution. *Nat. Commun*. **9**, 4968 (2018).

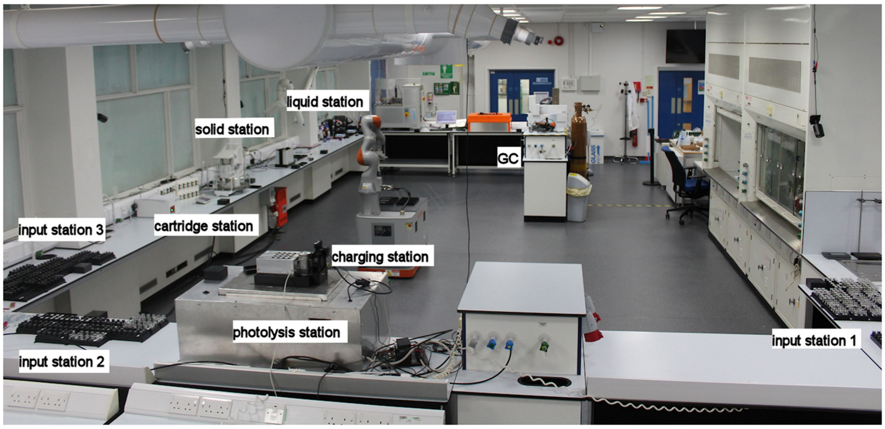
Acknowledgements We acknowledge financial support from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design, the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/N004884/1), the Newton Fund (grant number EP/R003580/1), and CSols Ltd. X.W. and Y.B. thank the China Scholarship Council for a PhD studentship. We thank KUKA Robotics for help with gripper design and the initial implementation of the robot.

Author contributions B.B. developed the workflow, developed and implemented the robot positioning approach, wrote the control software, designed the bespoke photocatalysis station and carried out experiments. P.M.M. and V.V.G. developed the optimizer and its interface to the control software. X.L. advised on the photocatalysis workflow. C.M.A., Y.B. and X.L. synthesized materials. Y.B. performed kinetic photocatalysis experiments. X.W. performed NMR analysis and synthesized materials. B.L. carried out initial scavenger screening. R.C. and N.R. helped to build the bespoke stations in the workflow. B.H. analysed the robustness of the system, assisted with the development of control software, and operated the workflow during some experiments. B.M.A. helped to supervise the automation work. R.S.S. helped to supervise the photocatalysis work. A.I.C. conceived the idea, set up the five hypotheses with B.B., and coordinated the research team. Data was interpreted by all authors and the manuscript was prepared by A.I.C., B.B., P.M.M., V.V.G. and R.S.S.

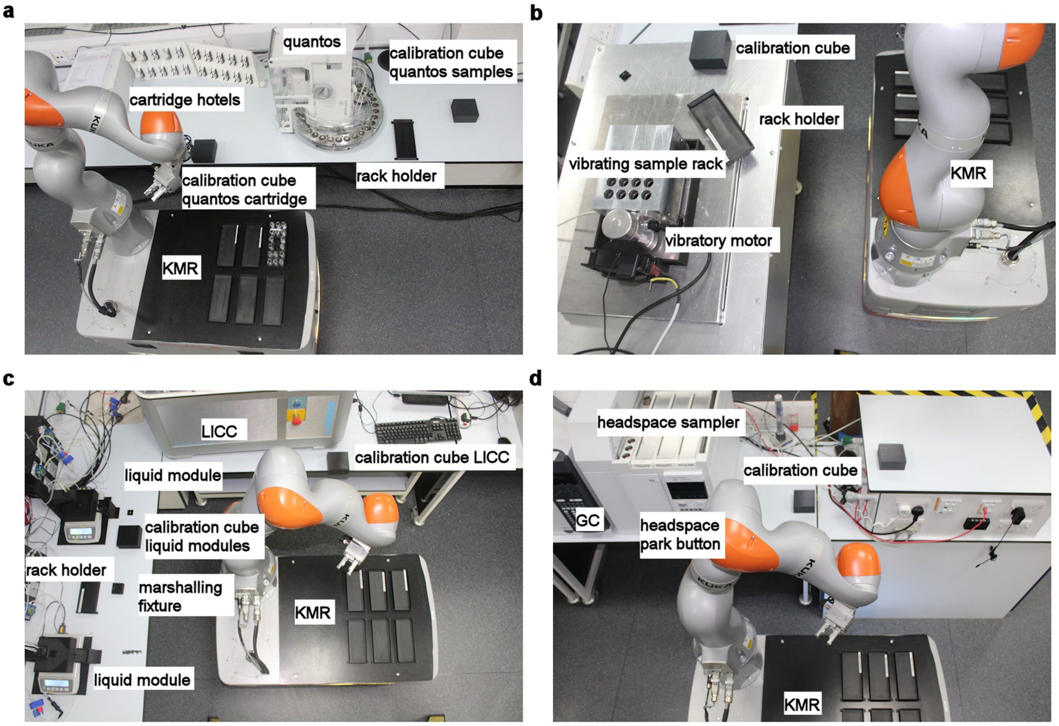
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2442-2.
Correspondence and requests for materials should be addressed to A.I.C.
Peer review information *Nature* thanks Volker Krueger, Tyler McQuade and Magda Titirici for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

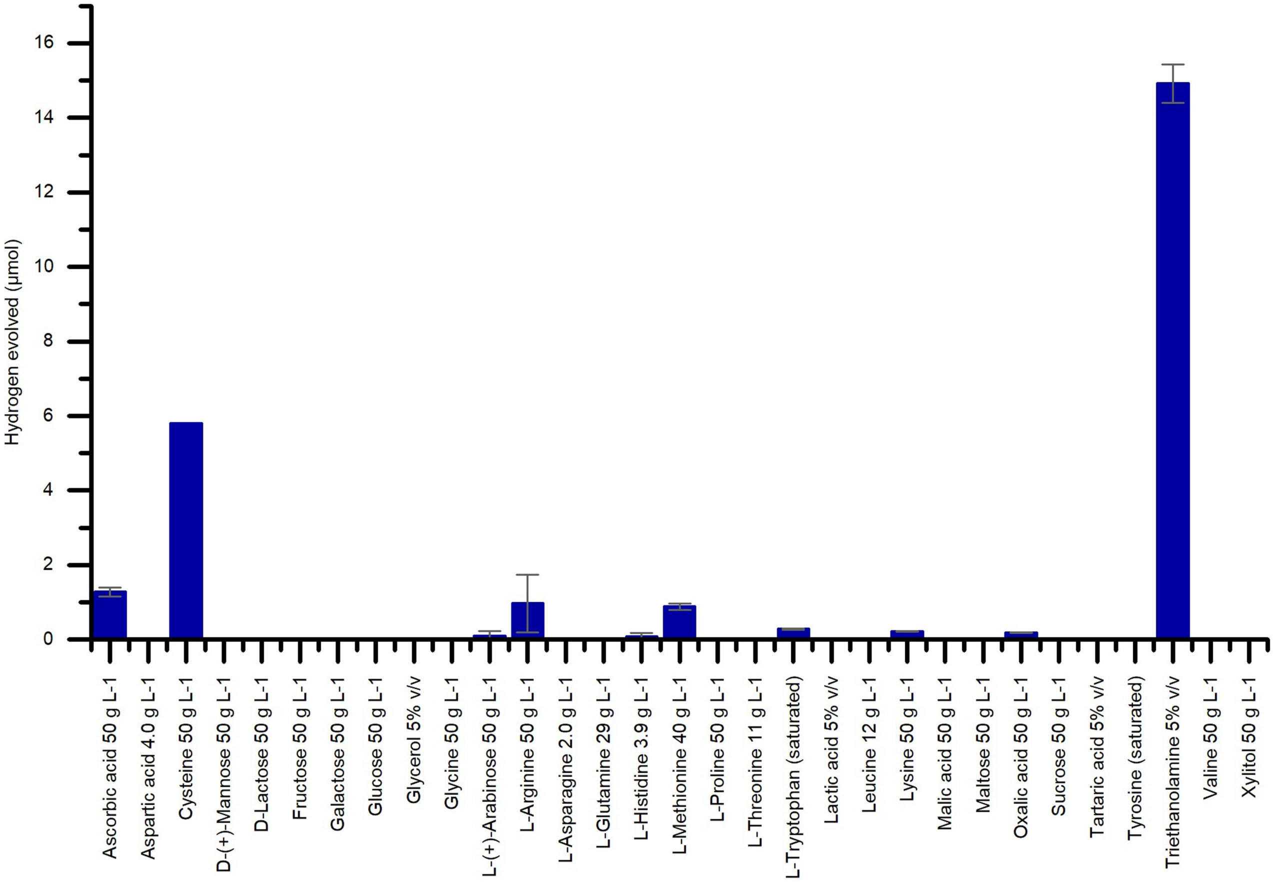
Extended Data Fig. 1 | Mobile robotic chemist. The mobile robot used for this project, shown here performing a six-point calibration with respect to the black location cube that is attached to the bench, in this case associated with the solid cartridge station (see also Supplementary Fig. 11 and Extended Data Fig. 3a).



Extended Data Fig. 2 | Laboratory space used for the autonomous experiments. The key locations in the workflow are labelled. Other than the black location cubes that are fixed to the benches to allow positioning (see also Extended Data Fig. 1), the laboratory is otherwise unmodified.

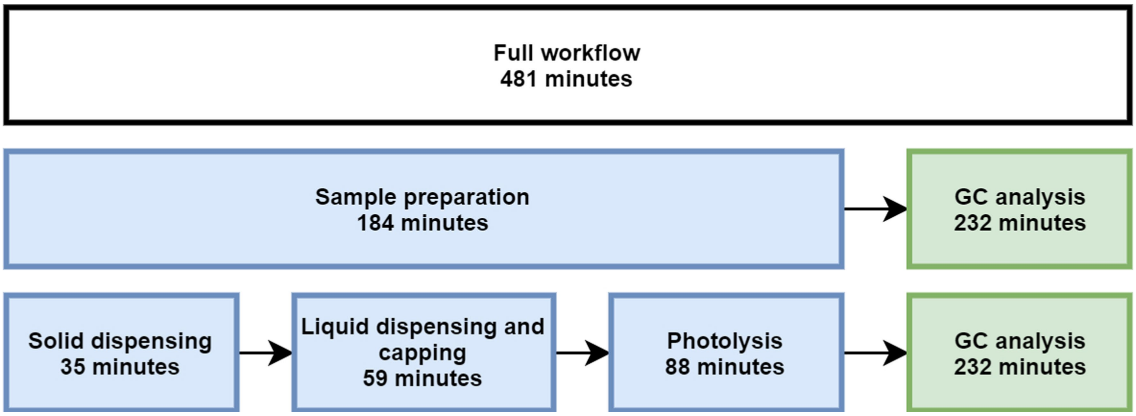


Extended Data Fig. 3 | Stations in the workflow. a, Photograph showing the robot at the solid dispensing / cartridge station. The two cartridge hotels can hold up to 20 different solids; here, four cartridges are located in the hotel on the left. The door of the Quantos dispenser is opened using custom workflow software that interfaces with the command software that is supplied with the instrument before loading the correct solid dispensing cartridge into the instrument (Supplementary Video 3). Since the KUKA Mobile Robot is free-roaming and has an 820 mm reach, it would be simple to extend this modular approach to hundreds or even thousands of different solids given sufficient laboratory space. **b**, Photograph showing the KUKA Mobile Robot at the photolysis station (see also Supplementary Videos 3, 6). **c**, Photograph showing the KUKA Mobile Robot at the combined liquid handling/capping station. The robot can reach both the liquid stations and the Liverpool Inertization Capper-Crimper (LICC) station after six-point positioning, such that liquid addition, headspace inertization and capping can be carried out in a single coordinated process (see Supplementary Videos 3, 5), without any position recalibration. **d**, Photograph of the KUKA Mobile Robot parked at the headspace gas chromatography (GC) station. The gas chromatography instrument is a standard commercial instrument and was unmodified in this workflow.



Extended Data Fig. 4 | Hydrogen evolution rates for candidate bioderived sacrificial hole scavengers. Results of a robotic screen for sacrificial hole scavengers using the mobile robot workflow. Of the 30 bioderived molecules trialed, only cysteine was found to compete with the petrochemical amine, triethanolamine. Scavengers are labelled with the concentration of the stock solution that was used (5 ml volume; 5 mg P10). The error bars show the standard deviation.

Extended Data Fig. 5 | Multipurpose gripper used in the workflow. The gripper is shown grasping various objects. **a**, The empty gripper; **b**, gripper holding a capped sample vial (top grasp); **c**, gripper holding an uncapped sample vial (side grasp); **d**, gripper holding a solid-dispensing cartridge; and **e**, gripper holding a full sample rack using an outwards grasp that locks into recesses in the rack. The same gripper was also used to activate the gas chromatography instrument using a physical button press (see Supplementary Video 3; 1 min 52 s).



Extended Data Fig. 6 | Timescales for steps in the workflow. Average timescales for the various steps in the workflow (sample preparation, photolysis and analysis) for a batch of 16 experiments. These averages were calculated over 46 separate batches. These average times include the time taken for the loading and unloading steps (for example, the photolysis time itself was 60 min; loading and unloading takes an average of 28 min per batch). The slowest step in the workflow is the gas chromatography analysis.

Potential for large-scale CO_2 removal via enhanced rock weathering with croplands

https://doi.org/10.1038/s41586-020-2448-9

Received: 31 May 2018

Accepted: 7 May 2020

David J. Beerling[1]✉, Euripides P. Kantzas[1], Mark R. Lomas[1], Peter Wade[1], Rafael M. Eufrasio[2], Phil Renforth[3], Binoy Sarkar[4], M. Grace Andrews[5], Rachael H. James[5], Christopher R. Pearce[6], Jean-Francois Mercure[7,8], Hector Pollitt[8,9], Philip B. Holden[10], Neil R. Edwards[8,10], Madhu Khanna[11], Lenny Koh[2], Shaun Quegan[12], Nick F. Pidgeon[13], Ivan A. Janssens[14], James Hansen[15] & Steven A. Banwart[16,17]

Enhanced silicate rock weathering (ERW), deployable with croplands, has potential use for atmospheric carbon dioxide (CO_2) removal (CDR), which is now necessary to mitigate anthropogenic climate change[1]. ERW also has possible co-benefits for improved food and soil security, and reduced ocean acidification[2–4]. Here we use an integrated performance modelling approach to make an initial techno-economic assessment for 2050, quantifying how CDR potential and costs vary among nations in relation to business-as-usual energy policies and policies consistent with limiting future warming to 2 degrees Celsius[5]. China, India, the USA and Brazil have great potential to help achieve average global CDR goals of 0.5 to 2 gigatonnes of carbon dioxide (CO_2) per year with extraction costs of approximately US\$80–180 per tonne of CO_2. These goals and costs are robust, regardless of future energy policies. Deployment within existing croplands offers opportunities to align agriculture and climate policy. However, success will depend upon overcoming political and social inertia to develop regulatory and incentive frameworks. We discuss the challenges and opportunities of ERW deployment, including the potential for excess industrial silicate materials (basalt mine overburden, concrete, and iron and steel slag) to obviate the need for new mining, as well as uncertainties in soil weathering rates and land–ocean transfer of weathered products.

The failure of the world to curb fossil fuel CO_2 emissions[6], and the inadequacy of planned mitigation measures[7], has been greeted with growing public consternation[8] consistent with the intergenerational injustice of human-caused climate change[9]. Even the most ambitious emission phase-outs[9,10] fail to achieve the United Nations Framework Convention on Climate Change Paris Agreement targets for limiting global warming without the help of massive amounts of atmospheric CDR. Extraction goals[1,7,9,10] later this century in most studies are on the order of at least 10 gigatonnes of CO_2 per year (Gt CO_2 yr^{-1}), although projections of rapid technological change[5] suggest a lower requirement of 2–2.5 Gt CO_2 yr^{-1}. This formidable challenge has led to international calls for urgent research into a portfolio of CDR options to understand their feasibility, scope, costs and challenges[11,12].

Our focus is terrestrial ERW, a CDR strategy based on amending soils with crushed calcium- and magnesium-rich silicate rocks to accelerate CO_2 sequestration[2–4,13–17]. Basalt, an abundant fast-weathering rock with the required mineral chemistry, could be ideal for implementing land-based ERW because of its potential co-benefits for crop production[18] and soil health[2–4]. ERW liberates base cations, generating alkalinity, so that atmospheric CO_2 is converted into dissolved inorganic carbon (principally hydrogen carbonate ions; HCO_3^-) that is removed via soil drainage waters. These weathering products are transported via land surface runoff to the oceans with a storage lifetime exceeding 100,000 years[19]. Depending on soil type and pH, atmospheric CO_2-derived dissolved inorganic carbon may also be sequestered through the formation of soil carbonate minerals, which reduces the efficiency of carbon sequestration by approximately half[19]. The logistical infrastructure to apply basaltic rock dust to managed croplands already exists owing to the common need to apply crushed limestone to reverse soil acidification resulting from intensive cropping[2–4]. Thus, rapid deployment at large scale appears to be feasible within decades, and has important ancillary benefits including mitigation of

---

[1]Leverhulme Centre for Climate Change Mitigation, Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK. [2]Advanced Resource Efficiency Centre, Management School, University of Sheffield, Sheffield, UK. [3]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK. [4]Lancaster Environment Centre, Lancaster University, Lancaster, UK. [5]School of Ocean and Earth Science, National Oceanography Centre Southampton, University of Southampton, Southampton, UK. [6]National Oceanography Centre, Southampton, UK. [7]Global Systems Institute, Department of Geography, University of Exeter, Exeter, UK. [8]Cambridge Centre for Energy, Environment and Natural Resource Governance, University of Cambridge, Cambridge, UK. [9]Cambridge Econometrics, Cambridge, UK. [10]Environment, Earth and Ecosystems, The Open University, Milton Keynes, UK. [11]Department of Agricultural and Consumer Economics, Institute for Sustainability, Energy, and Environment, University of Illinois, Urbana, IL, USA. [12]School of Mathematics and Statistics, Hicks Building, University of Sheffield, Sheffield, UK. [13]Understanding Risk Research Group, School of Psychology, Cardiff University, Cardiff, UK. [14]Research Group Plants and Ecosystems, University of Antwerp, Antwerp, Belgium. [15]Earth Institute, Columbia University, New York, NY, USA. [16]Global Food and Environment Institute, University of Leeds, Leeds, UK. [17]School of Earth and Environment, University of Leeds, Leeds, UK. ✉e-mail: d.j.beerling@sheffield.ac.uk

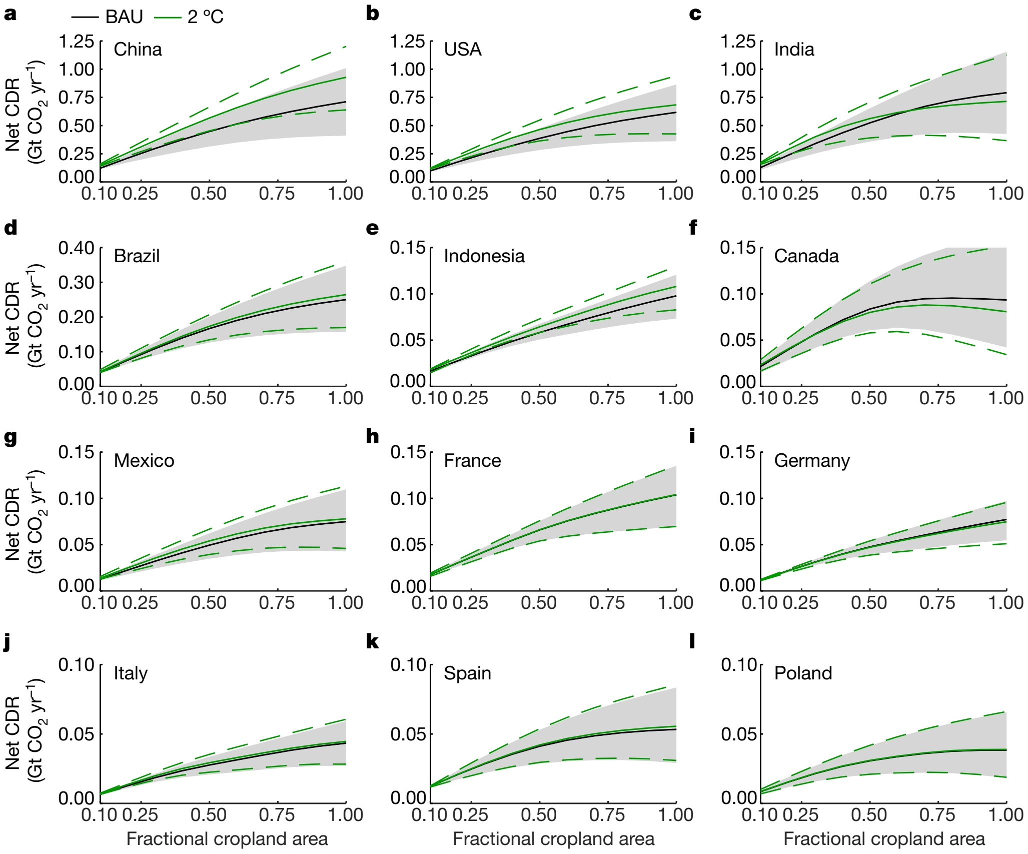


Fig. 1 | CDR via ERW with croplands. Net CDR curves for nations with the highest CDR potential worldwide (**a**–**g**) and in Europe (**h**–**l**) as a function of increasing ERW deployment across existing croplands. Note the *y*-axis scale changes. Results are shown for the BAU and the 2 °C energy policy scenarios. The grey-shaded area for each nation represents the 90% confidence interval calculated for basalts with relatively slow- versus fast-weathering rates for the BAU scenario; short green dashed lines indicate the 90% confidence limits of the corresponding 2 °C scenario simulations. Uncertainty in net CDR increases as ERW deploys onto croplands occupying a wider range of environmental conditions.

ocean acidification[15–18]. Carbon sequestration by ERW on croplands—a biogeochemical CDR option supporting multiple United Nations sustainable development goals and ecosystem services[4,20], and a pragmatic land-use choice to maximize scalability and co-benefits—thus warrants detailed examination.

We constructed a performance model with sub-national level of detail to assess quantitatively the CDR capacity and costs for land-based ERW implementation in major economies, constrained by available agricultural land area and energy production (including USA, India, China, Brazil and Europe) (Extended Data Fig. 1). For rock weathering within the soil profile, we developed a one-dimensional vertical reactive transport model with steady-state flow, and a source term representing rock grain dissolution (Methods; Supplementary Figs. 1–12; Supplementary Tables 1–5). Our work builds on advances made in prior ERW research largely on tropical forested ecosystems[15–17,21,22], with the practical aims of understanding the capacity of agriculture to capture carbon via soil amendment with milled basalt. For this initial nation-by-nation assessment, we examine the sensitivity of net CDR on current croplands to the projected national energy production for 2050 under a business-as-usual (BAU) energy scenario based on ongoing energy transitions[5]. This is compared with a 2 °C scenario (that is, a scenario in which the increase in global mean temperature since the pre-industrial period is limited to 2 °C by 2050), which includes a wide range of policy measures designed to respect the 2 °C target with 75% probability[5] (Supplementary Tables 6–12).

CDR potential via ERW

Our geospatial analyses define a new technical potential CDR range for those nations with high capacity for ERW deployment on cropland (Fig. 1; Supplementary Figs. 13–15). For each nation, we generate CO_2 capture curves by ranking CDR potential from the highest to the lowest grid cells with increasing ERW deployment. National median CDR curves typically show CDR capacity rising with increasing cropland area, with CDR by silicate soil amendment reaching a plateau, or declining in the case of Canada (Fig. 1). These patterns reflect expansion of ERW into climatically unfavourable agricultural land, causing CDR potential to slow relative to the carbon penalty of logistical operations, and our assumed 3% limit in national energy available for grinding (see Methods section 'Cost assessments' and Extended Data Fig. 2). Overall trends in national CDR curves are relatively insensitive to the choice of energy scenario. China is the exception because its large increase in low-carbon energy usage projected under the 2 °C scenario[5] allows net CDR to rise by substantially reducing secondary CO_2 emissions from logistical operations (Fig. 1). This contrasts with results for India, whose total energy production falls by around 40% with a transition to low-carbon energy production in the 2 °C scenario, lowering the energy available for grinding basalt, and thus the potential for increased CDR by ERW. Reductions in energy production for other nations in the 2 °C scenario compared with the BAU scenario similarly lower their potential for increased CDR with the transition to low-carbon energy.

Recognizing the urgent need to assess large-scale options for meeting near-term CDR goals[10], we determine the potential contribution of nations to achieve CDR goals across the 0.5–2 Gt $CO_2\,yr^{-1}$ range (Table 1; Extended Data Fig. 3). Overall, we find that the three countries with the highest CDR potential are coincidentally the highest fossil fuel CO_2 emitters (China, USA and India)[6] (Fig. 1). Indonesia and Brazil, with CO_2 emissions 10–20 times lower than the USA and China, have relatively high CDR potential owing to their extensive agricultural lands and warm, seasonally wet climates conducive to high silicate rock weathering efficiency. European countries have a CDR potential an order of magnitude lower than those of China, USA and India, mainly because of less agricultural land area. The five European nations with the highest net CDR potential could offset 30% of the current emissions of European nations in 2019 and the three European countries with the highest CDR potential are also the largest European emitters of CO_2 from fossil fuels (Germany, Spain and Poland)[6]. Our ERW scenarios (Table 1) correspond to an aggregate CDR of 25–100 Gt CO_2

if sustained over five decades. This would save up to 12% of the remaining cumulative carbon emission budget (about 800 Gt CO_2) that gives a 66% probability of limiting global warming to below 2 °C above the pre-industrial average surface temperature[10].

In the context of our CDR goals, ERW has a potential similar to that of other CDR strategies[23] estimated for 2050, including bio-energy with carbon capture and storage (BECCS), widely adopted in IPCC future scenarios (0.5–5 Gt CO_2 yr^{-1}), direct air capture and storage (DAC) (0.5–5 Gt CO_2 yr^{-1}), biochar (0.5–2 Gt CO_2 yr^{-1}), soil organic carbon sequestration (0.5–5 Gt CO_2 yr^{-1}), and afforestation/reforestation (0.5–3.6 Gt CO_2 yr^{-1}). One benefit of country-level analysis for CDR is the scope for comparative assessments with other technologies and opportunities for co-deployment. For example, our ERW CDR range is comparable with large-scale implementation of BECCS in the USA by 2040 (0.3–0.6 Gt CO_2 yr^{-1}), as constrained by biomass productivity, location and capacity of CO_2 storage sites[24]. ERW avoids competition for land used in food production, and related increased demands of BECCS for freshwater and polluting fertilizers[25], with CO_2 being treated as a resource for mineral weathering. Co-deployment of ERW with feedstock crops for BECCS and biochar could enhance the feasibility and carbon sequestration potential of these strategies[4,26].

Inorganic carbon sequestration by ERW appears to be comparable to soil organic carbon sequestration, another proposed CDR strategy (about 2.5 Gt CO_2 yr^{-1} by 2100)[27] using agricultural land, but with potentially greater long-term security of carbon storage. Co-deployment of ERW and soil organic carbon sequestration at large scale might, therefore, contribute substantially to the 5 Gt CO_2 yr^{-1} CDR goal suggested in decarbonization scenarios[10] for 2050. Compatibility of ERW and soil organic carbon sequestration may be realistic given that amendment of acidic organic-rich soils with silicate minerals, and the resultant pH increase, had no effect on respiratory CO_2 fluxes[28,29], contrary to concerns that increased soil pH may accelerate organic matter decomposition[30]. However, efficacy of CDR, sink saturation, and permanency of storage with these approaches, separately and interactively, are uncertain[11,23]. Abatement of soil N_2O emissions by basalt application to conventionally managed arable and perennial crops[31], and of N_2O and CH_4 emissions by application of artificial silicates to rice agriculture[32], is possible. Such effects would further lower adverse impacts of agriculture on climate per unit yield, amplifying the climate mitigation potential of ERW.

Greenhouse gas emissions reductions aimed at limiting future warming are defined under the Paris Agreement by Nationally Determined Contributions (NDCs)[33]. As yet, most of the top ten fossil carbon emitting nations are failing to meet their 2030 NDC pledges which, even if met, imply a median global warming (2.6–3.1 °C) exceeding the Paris agreement[33]. Warming of this magnitude could allow the Earth system to cross thresholds for irreversible planetary heating and long-term multi-metre sea-level rise, with potentially disastrous consequences for coastal cities[34]. NDC pledged carbon emission reductions undergo periodic revision in response to trends in greenhouse gas emissions, uptake of low-carbon energy technology, and climate[33] and hence are not set for 2050. We therefore illustrate the potential for undertaking ERW with agricultural lands to strengthen near-term national 2030 NDCs (Fig. 2).

Table 1 | CDR goals for ERW with croplands in 2050

Goal		Cropland area (%)	National CDR (Gt CO_2 yr^{-1})	Silicate demand (Gt yr^{-1})	Cost (US$ per t CO_2 yr^{-1})
0.5 Gt CO_2 yr^{-1}					
World	China	10	0.13	0.77	102.1
	USA	11	0.11	0.63	160.3
	India	11	0.15	0.84	78.4
	Brazil	10	0.041	0.22	123.8
	Indonesia	10	0.017	0.091	54.3
	Canada	10	0.022	0.13	177.6
	Mexico	10	0.013	0.073	97.5
Europe	France	10	0.017	0.085	158.1
	Germany	11	0.012	0.066	167.8
	Italy	11	0.0070	0.039	181.9
	Spain	10	0.012	0.066	192.8
	Poland	10	0.0085	0.050	171.6
1.0 Gt CO_2 yr^{-1}					
World	China	23	0.26	1.59	109.3
	USA	24	0.21	1.26	168.5
	India	23	0.24	1.50	79.9
	Brazil	23	0.083	0.45	116.4
	Indonesia	25	0.033	0.18	57.5
	Canada	16	0.030	0.20	191.7
	Mexico	23	0.025	0.15	103.1
Europe	France	24	0.034	0.17	160.4
	Germany	25	0.025	0.14	171.7
	Italy	23	0.014	0.083	191.0
	Spain	17	0.018	0.10	190.9
	Poland	17	0.012	0.081	170.9
1.5 Gt CO_2 yr^{-1}					
World	China	38	0.40	2.48	114.5
	USA	39	0.32	1.99	173.1
	India	36	0.37	2.35	80.2
	Brazil	36	0.13	0.71	110.5
	Indonesia	41	0.050	0.28	58.6
	Canada	25	0.045	0.35	207.3
	Mexico	37	0.038	0.23	105.6
Europe	France	38	0.050	0.26	159.5
	Germany	39	0.037	0.20	173.6
	Italy	37	0.021	0.13	194.1
	Spain	28	0.026	0.17	189.3
	Poland	27	0.019	0.13	171.3
2.0 Gt CO_2 yr^{-1}					
World	China	55	0.53	3.46	120.7
	USA	55	0.42	2.72	176.7
	India	51	0.49	3.30	80.9
	Brazil	51	0.17	0.98	106.2
	Indonesia	59	0.067	0.38	59.4
	Canada	35	0.060	0.51	220.3
	Mexico	52	0.050	0.33	106.8
Europe	France	54	0.067	0.36	157.1
	Germany	57	0.050	0.28	175.9
	Italy	55	0.029	0.18	193.3
	Spain	41	0.035	0.25	190.7
	Poland	38	0.025	0.19	175.4

All values are means of both energy scenarios; see main text for details. For each country, *c*, we assigned its contribution, CDR(c), to a global CDR goal as follows, where $CDR_{max}(c)$ is the maximum CDR value attainable by a country: $CDR(c) = CDR_{goal}\frac{CDR_{max}(c)}{\sum_{countries} CDR_{max}(c)}$.

Results show that China may be able to augment its pledged 2030 NDCs by about 5% to 10%, with similar gains for the USA, which has opted out of the Paris agreement. For India, the gain rises to 40% of its current pledged emissions, and Brazil may be able to offset 100% of its pledged 2030 CO_2 emissions plus some fraction of those from other countries (Fig. 2). Other countries outside Europe considered in our analysis (Indonesia, Canada, Mexico) may be able to augment their NDCs by up to 30% (Fig. 2). In Europe, ERW could aid substantial decarbonization of France and Spain (up to approximately 40%), and to a lesser extent Poland, Italy and Germany (all about 10%) (Fig. 2). ERW, therefore, may have a role to play in compensating for residual carbon

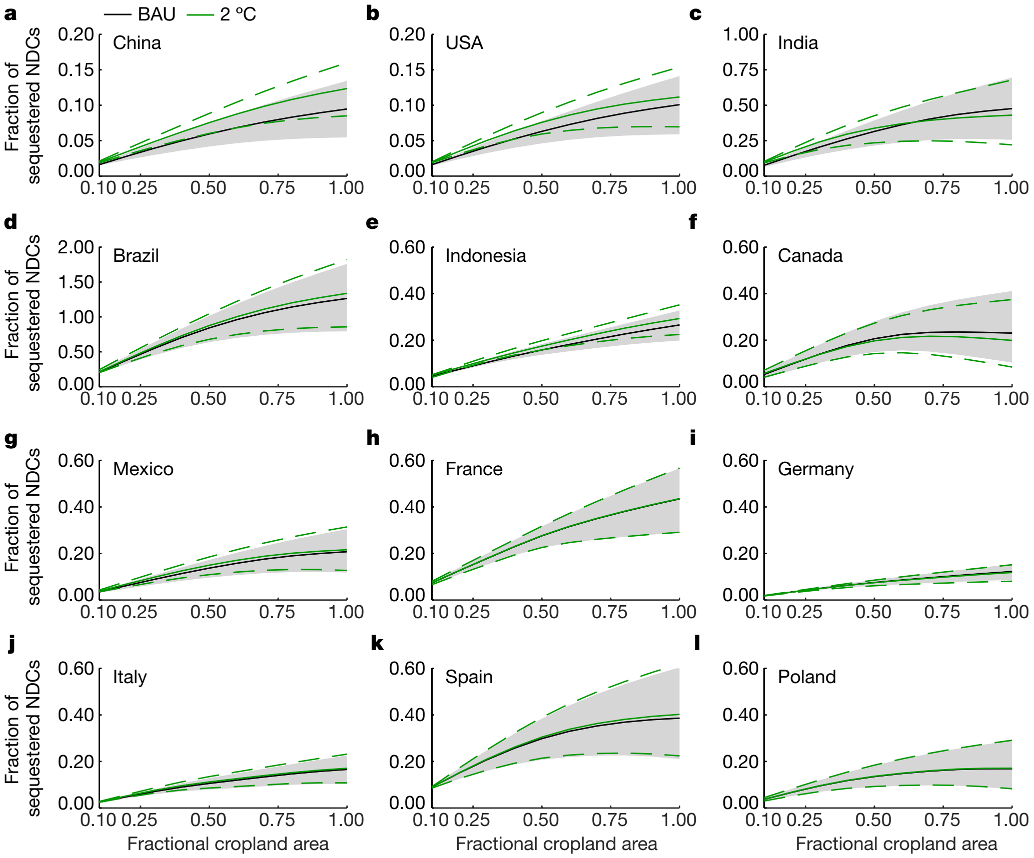


Fig. 2 | Augmentation of pledged CO_2 emissions reduction by ERW. Fraction of 2030 NDC emissions reductions by enhanced weathering for nations with the highest CDR potential worldwide (**a**–**g**) and in Europe (**h**–**l**), as a function of increasing ERW deployment across croplands. Note the *y*-axis scale changes. Results are shown for the BAU energy policy and the 2 °C energy policy scenarios. The grey-shaded area for each nation represents the 90% confidence interval calculated for basalts with relatively slow- versus fast-weathering rates for the BAU scenario; short green dashed lines indicate the 90% confidence limits of the corresponding 2 °C scenario simulations.

emissions from sectors recognized as being difficult to decarbonize, for example, transportation by aviation, shipping and agriculture[1,11].

Costs of CDR via ERW

Cost assessment is needed to evaluate commercial feasibility of ERW and to put a price on climate mitigation actions (Extended Data Fig. 4). Our cost estimates based on current prices (2019 US dollars, throughout) fall within the range of prior ERW assessments (US$75–250 per tonne of CO_2)[21–23] while resolving differences among nations (Fig. 3; Table 1; Supplementary Figs. 16–25; Supplementary Tables 13 and 14). Average costs in USA (US$160–180 per tonne of CO_2), Canada and European nations (US$160–190 per tonne of CO_2) are almost 50% higher than those in China, India, Mexico, Indonesia and Brazil (US$55–120 per tonne of CO_2). The difference largely reflects labour, diesel and electricity costs.

Defined as the cost of CDR and storage, the price of carbon is a proposed economic enabler for bringing CDR strategies to market[11]. Carbon price is forecast by the World Bank[11] to reach US$100–150 per tonne of CO_2 by 2050. Costs per tonne of CO_2 removed by ERW are generally within this projected carbon price range in all nations, but unit costs increase when cropland area exceeds the optimal fraction, because the efficiency of weathering and CDR falls (Fig. 3; Table 1). A carbon price of US$100–150 per tonne of CO_2 would cover most of the ERW costs for the key nations reported here. It would make ERW an economically attractive option for fast-growing nations, such as India, China, Indonesia, Brazil and Mexico, given their estimated CO_2 extraction costs of around US$75–100 per tonne of CO_2 (Fig. 3; Table 1).

Our estimated ERW costs of CDR for nations are comparable to estimates summarized for BECCS (US$100–200 per tonne of CO_2), direct air capture and storage (US$100–300 per tonne of CO_2), and biochar (US$30–US$120 per tonne of CO_2), but higher than estimates for soil organic carbon sequestration (US$0–10 per tonne of CO_2)[23]. Afforestation/reforestation and practices that increase soil carbon in natural ecosystems, including wetland restoration, have lower estimated costs (<US$100 per tonne of CO_2)[23]. However, these natural carbon sequestration options require assessment for possible indirect unintended positive climate feedbacks.

Per capita metrics help to conceptualize the matter of costs in terms relevant to citizens. Current fossil fuel emissions per person per year[6] are 16.5 t CO_2 (USA), 15.1 t CO_2 (Canada), 7.5 t CO_2 (China), 7.3 t CO_2 (Europe), 2.6 t CO_2 (Brazil), 1.8 t CO_2 (Indonesia) and 1.7 t CO_2 (India). ERW cannot offset all fossil fuel emissions, but using its cost as a guide, the per capita annual cost of achieving zero net emissions, a goal for decarbonization, would be highest for Canada (US$3,004), the USA (US$2,780), China (US$832) and Europe (US$1,288). Costs fall substantially for citizens in Brazil (US$300), Indonesia (US$103) and India (US$135) (Table 1).

At this early stage of research and development, costs are uncertain and ERW is in need of demonstration projects[7,11,12]. Costs are likely to decline as the market expands and technologies develop. This includes emergence of more energy-efficient, low-carbon technologies for rock grinding. Costs may also decline via co-deployment with afforestation/reforestation projects or agroforestry as part of worldwide carbon-offset trading schemes[7]. The net cost of ERW may be lower, given that rock dust is an acceptable fertilizer for organic agriculture, which currently occupies 57.8 million hectares, because it adds economic value by improving soil health, fertility and ecosystem services[35].

Implementation challenges and opportunities

Our analysis of the techno-economic potential for CDR via ERW strengthens the case for evaluating all aspects of practical deployment in developed and developing economies. This includes: meeting rock demand through alternative sources that avoid mining expansion;

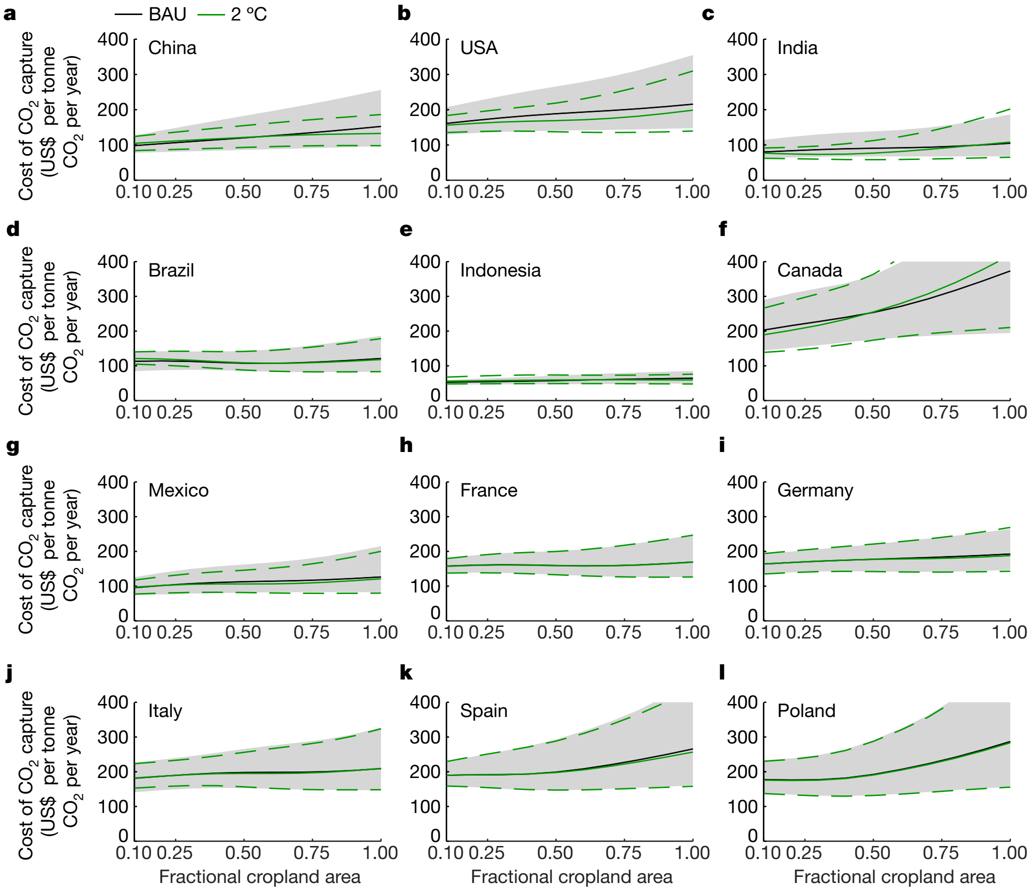


Fig. 3 | Costs of carbon extraction via ERW with croplands. Costs of CO_2 extraction from air by ERW for nations with the highest CDR potential worldwide (**a**–**g**) and in Europe (**h**–**l**), as a function of increasing ERW deployment across croplands. Results are shown for the BAU and the 2 °C energy policy scenarios. The grey-shaded area for each nation represents the 90% confidence interval calculated for basalts with relatively slow- versus fast-weathering rates for the BAU scenario; short green dashed lines indicate the 90% confidence limits of the corresponding 2 °C scenario simulations.

undertaking a more complete economic valuation; and engaging the public to understand social acceptance.

National demand for crushed silicate rock is dependent on the extent of ERW deployment (Extended Data Fig. 5). Within our scenarios, the demand for basalt required for ERW rises with an increasing CDR goal and scales with agricultural land area (Table 1). Safeguarding against substantial increased mining and possible adverse impacts on inhabitants[36] requires exploiting underutilized stockpiles of crushed basalt produced as a by-product of the aggregate industry. Mining generates a continuous, but often discarded, finely powdered silicate by-product that is utilizable for ERW. These materials have been accumulating worldwide for decades and may require no further grinding, thus lowering CO_2 emissions that reduce CDR efficiency (Extended Data Fig. 6)[21–23]. However, national inventories of the location, availability and extent of this resource are required to assess the potential contribution of this resource to CDR via ERW.

Requirement for mining may be further reduced by using artificial silicate by-products from industrial processes[37,38], including calcium-rich silicates produced by iron and steel manufacturing (slags) with a long history of agricultural usage[4,39]. This material is recycled as low-value aggregate (less than about US$5 per tonne), and often stockpiled at production sites or disposed of in landfills, whereas it could become a valuable commodity for CDR. The largest amounts of by-products from the construction and demolition industry are cement, sand and masonry. Following separation from other materials (for example, metals and plastics), the cement comprises relatively 'clean' calcium-rich silicates and may be suitable for application to soils, but this suggestion requires field trials to assess suitability. Cement contributes about 6% to global CO_2 emissions[6] and ERW may represent a land management option for valuing the by-products of cement and improving the sustainability of this worldwide industry.

We forecast production of artificial calcium-rich cements for construction and by-product slag from steel manufacturing for Brazil, China, India and the USA, to understand their potential role in meeting silicate demand for ERW (Fig. 4). Differences between national production estimates are driven by forecast population increases over the coming century, and per capita consumption trends for the material under the middle-of-the-road Shared Socioeconomic Pathway (see Methods). Bulk silicate production from the construction and demolition sector is modelled to increase substantially in all four nations, with China and India having a combined production by 2060 of about 13 Gt yr^{-1} (Fig. 4). China and India dominate, with above-average per capita cement use compared to the global average, and substantially larger populations than the USA and Brazil[38]. Thus, bulk silicate production of these two nations could meet the demand for ERW with large CDR potential (Table 1). Although chemically similar to basalt, these artificial calcium-rich silicates contain minerals that dissolve several orders of magnitude faster, react rapidly with CO_2 in soils under ambient conditions[40], and are produced in fine particle sizes that facilitate accelerated weathering[41].

Agricultural production could benefit substantially from increased resource use efficiency, reducing consumption of raw materials and recovering mineral nutrients from silicate by-products[32,42,43] and from legacy reserves of silicate rock dust[44]. However, application of any silicate material to agricultural soils requires careful assessment of the risks, including potential release of metals and persistent organic compounds (Supplementary Table 15). Undertaking ERW practices with these materials addresses a critical need to fertilize soils with silica and other nutrients lost by harvesting that gradually depletes plant-available pools[39]. Intensification of food production across 24 million hectares of productive agricultural land in South Asia and China, for example, is creating acidified, desilicated soils exhausted in plant nutrients (potassium, zinc and available phosphorus) that limit yields[45]. Yet these negative effects may be reversible with ERW treatments such as fertilization of irrigated rice using either natural and/or artificial silicates (for example, recycled steel slags). Such treatments replenish plant available silica pools, increasing yields and soil pH, and decrease the mobility of potentially toxic trace elements (for example, arsenic)[46]. ERW may therefore also have a role in remediation of toxic-metal-contaminated soils and sediments across 20 million hectares of cultivated land in southern China and elsewhere[47].

More broadly, innovative ERW practices via soil amendments with targeted silicate minerals could help to rebuild rapidly deteriorating agricultural soils on which over six billion people depend directly for food[48]. Such practices may complement other approaches to soil improvement, including conservation tillage and nitrogen-fixing cover crops. The current substantial rate of agricultural top-soil depletion requires urgent remedial action, with high economic costs apparent already in China, where degradation of soils supporting wheat, maize and rice production costs an estimated US$12 billion annually[48]. Targeted amendment of agricultural soils for CDR may have a role in slowing rates of soil loss by up to 45%, with the accelerated weathering of added minerals replacing inorganic nutrients and the resultant formation of clays and mineral organic aggregates increasing the cation exchange capacity and water storage capacity of rebuilt soils[4,20]. The addition of trace amounts of zinc and iron could also improve public health by reversing the effect of rising CO_2 levels on the declining nutritional value of food crops[49].

The feasibility of mobilizing millions of smallholder communities to adopt ERW practices in China and India will depend both on demonstrating that soil improvements can reverse yield declines and on government subsidies. Farming practices adopted for increasing sustainable productivity, for example, have transformed agriculture across 37 million hectares in China, increasing profits by US$12.2 billion over a decade[50]. With 2.5 billion smallholders farming 60% of the world's arable land, a similar outreach programme could be used throughout Asia, with farmers earning more profits from higher yields while sequestering CO_2. Involving local scientists in conducting research into its effectiveness and safety to build trust and engagement with smallholder farmers is key, alongside involvement with policymakers and stakeholders. This increases the potential to bring smallholders out of extreme poverty and, in the regions with climates suitable for non-irrigated agriculture, restore highly degraded soils not suitable currently for food production.

Realizing the potential of ERW as a biogeochemical approach to sequester CO_2 by altering land management practices will depend on the commitment of farmers and governments, implementation of the right policy frameworks and wider public acceptance. Understanding the balance between positive and negative outcomes, in terms of public acceptance of the inevitable trade-offs between local mining activities versus global sequestered carbon, requires empirical testing with stakeholders and the wider public. Crucially, such testing needs to understand the conditions that society might place upon the development and large-scale deployment of ERW technologies, as part of a wider responsible research and innovation programme[51].

Uncertainties

Our analysis of the techno-economic potential of CDR by ERW is subject to several uncertainties, particularly variation in our baseline application rate and basalt mineralogy. It also identifies priority areas benefitting from more research into ERW under field conditions.

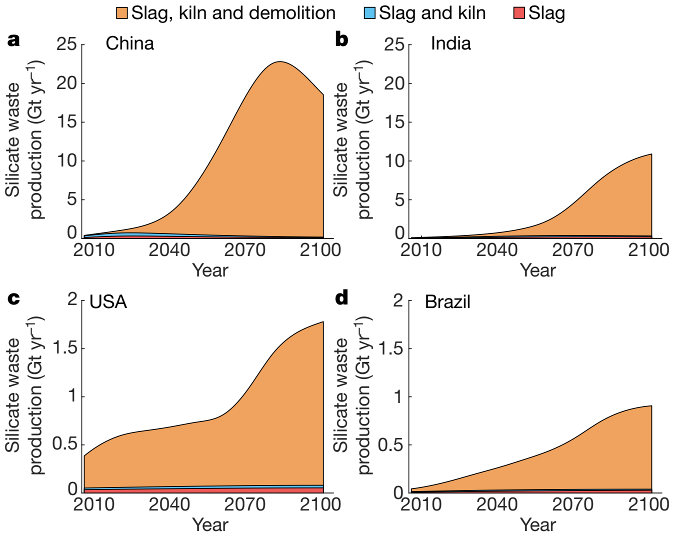


Fig. 4 | Forecast increases in national bulk silicate production over the next century. Simulated future increases in bulk artificial silicate by products (slag, cement, kiln dust and cementitious demolition waste) production during the twenty-first century are given for China (**a**), India (**b**), USA (**c**) and Brazil (**d**), based on the middle-of-the-road Shared Socioeconomic Pathway (Methods).

Extrapolation of laboratory weathering rates to the field scale is a recognized potential source of uncertainty in calculated CDR rates by ERW[2–4,22–24]. We addressed this by Monte Carlo analysis of the fractal dimension accounting for uncertainty in the apparent reacting surface area of grains for ERW conducted at large geographical scales. Together with the chemical affinity effects accounted for in our model, we constrain some of the systematic errors embedded in prior ERW assessments[15–17,21,22].

Surface passivation, a component of chemical inhibition, occurs as weathering proceeds, creating leached layers and relatively stable secondary minerals, which potentially inhibit the mass transfer kinetics of elements from the dissolving surfaces of primary minerals. The current state of knowledge[52] precludes a detailed treatment of the role of surface passivation by formation of amorphous silica-rich surfaces for basalt grains added to agricultural soils. ERW analysis will benefit from future research to improve mechanistic insight and formulation of kinetic equations.

It remains to be determined whether our theoretical analyses of the techno-economic potential for this CDR approach are consistent with findings from long-term field-scale ERW trials. Such trials are urgently required to assess weathering and CDR efficiency of freshly crushed rock grains with highly reactive surfaces added to agricultural soils subject to periodic wet–dry cycles during the growing season[3]. The potential for the trapping of weathered cations on ion exchange surfaces or within secondary minerals other than carbonates, thus delaying or preventing land–ocean transfer, will depend on soil type, climate, hydrological conditions, application rate and management practices. The duration of the carbon sequestration rate, and the possibility of CO_2 sink saturation with ERW on croplands, are both poorly constrained by data, as for other land-based CDR strategies[11,23]. Other areas for further research include: the quantification of biogeochemical transformations of carbon and nitrogen associated with organic and inorganic fertilization practices; atmospheric deposition; and the role of rhizosphere biology.

Conclusions

The techno-economic assessment of ERW's potential to contribute large-scale CDR requires further integration of nation-by-nation quantitative analysis, together with large-scale pilot demonstrations supported by fundamental process studies and public engagement. Our analysis identifies the engineering challenges if ERW were to be scaled up to help meet ambitious CDR goals as part of a wider portfolio of options[1,7,11,12]. ERW estimated costs are comparable to (and generally lower than) current estimates for the intensive CDR technologies—BECCS and direct air capture—and have potential ancillary benefits through limiting coastal zone acidification and improving food and soil security. Nations that may have large ERW potential, including China, the USA and India, are all vulnerable to climate change and resultant sea-level rise[34]. Their high risks of economic damage[53] and social disruption should provide the impetus for creative co-design of agricultural and climate policies. Success requires incentives and regulatory frameworks that overcome social and political inertia. Silicate demand of nations must also be met in a way that facilitates social acceptance[51,54] and preservation of biodiversity[4,20].

Deployment of any CDR strategy is inhibited by the concern that it may erode society's perception of the climate threat and the urgency of mitigation measures[54]. However, the ancillary benefits of ERW may aid its early use by creating 'demand pull', and relieve such concern. Innovative 'climate-smart' farming practices can be designed with ERW to draw down CO_2 and other greenhouse gases while recycling nutrients, aiding soil water storage, and supporting crop production[4,18,20]. Such practices can help to restore deteriorating topsoils that underpin food security for billions of people while maximizing the societal co-benefits needed to incentivize deployment[20]. Financial, industrial and policy road-mapping that links the possibility to reliably set and achieve short-term and long-term goals is needed, including a broader analysis of risks[23] and co-benefits[2–4,18,20], to determine the part that ERW might play in climate risk mitigation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2448-9.

1. Intergovernmental Panel on Climate Change (IPCC). *Global Warming Of 1.5 °C. An IPCC Special Report on the Impacts of Global Warming of 1.5 °C Above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways* (World Meteorological Organization, 2018).
2. Kantola, I. B. et al. Potential of global croplands and bioenergy crops for climate change mitigation through deployment for enhanced weathering. *Biol. Lett*. **13**, 20160714 (2017).
3. Zhang, G., Kang, J., Wang, T. & Zhu, C. Review and outlook for agromineral research in agriculture and climate change mitigation. *Soil Res*. **56**, 113–122 (2018).
4. Beerling, D. J. et al. Farming with crops and rocks to address global climate, food and soil security. *Nat. Plants* **4**, 138–147 (2018).
5. Mercure, J.-F. et al. Macroeconomic impact of stranded fossil fuel assests. *Nat. Clim. Chang*. **8**, 588–593 (2018).
6. Le Quéré, C. et al. Global carbon budget 2018. Earth Syst. *Sci. Data* **10**, 2141–2194 (2018).
7. United Nations Environment Programme *The Emissions Gap Report 2018* (United Nations Environment Programme, 2018).
8. Hagedorn, G. et al. Concerns of young protesters are justified. *Science* **364**, 139–140 (2019).
9. Hansen, J. et al. Young people's burden: requirement of negative CO_2 emissions. *Earth Syst. Dyn* **8**, 577–616 (2017).
10. Rockström, J. et al. A roadmap for rapid decarbonisation. *Science* **355**, 1269–1271 (2017).
11. *The Royal Society Greenhouse Gas Removal Technologies* (The Royal Society, 2018).
12. Pacala, S. et al. *Negative Emissions Technologies And Reliable Sequestration* (National Academy of Sciences, 2018).
13. Seifritz, W. CO_2 disposal by means of silicates. *Nature* **345**, 486 (1990).
14. Schuiling, R. D. & Krijgsman, P. Enhanced weathering: an effective and cheap tool to sequester CO_2. *Clim. Change* **74**, 349–354 (2006).
15. Kohler, P., Hartmann, J. & Wolf-Gladrow, D. A. Geoengineering potential of artificially enhanced silicate weathering of olivine. *Proc. Natl Acad. Sci. USA* **107**, 20228–20233 (2010).
16. Hartmann, J. et al. Enhanced chemical weathering as a geoengineering strategy to reduce atmospheric carbon dioxide, supply nutrients, and mitigate ocean acidification. *Rev. Geophys*. **51**, 113–149 (2013).
17. Taylor, L. L. et al. Enhanced weathering strategies for stabilizing climate and averting ocean acidification. *Nat. Clim. Chang*. **6**, 402–406 (2016).
18. Kelland, M. E. et al. Increased yield and CO_2 sequestration potential with the C_4 cereal crop *Sorghum bicolor* cultivated in basaltic rock dust amended agricultural soil. *Glob. Change Biol*. **26**, 3658–3676 (2020).
19. Renforth, P. & Henderson, G. Assessing ocean alkalinity for carbon sequestration. *Rev. Geophys*. **55**, 636–674 (2017).
20. Smith, P. et al. Land-based options for greenhouse gas removal and their impacts on ecosystem services and the sustainable development goals. Annu. *Rev. Environ. Res*. **44**, 255–286 (2019).
21. Renforth, P. The potential of enhanced weathering in the UK. *Int. J. Greenhouse Gas Control* **10**, 229–243 (2012).
22. Strefler, J. et al. Potential and costs of carbon dioxide removal by enhanced weathering of rocks. *Environ. Res. Lett*. **13**, 034010 (2018).
23. Fuss, S. et al. Negative emissions—Part 2: Costs, potentials and side effects. *Environ. Res. Lett*. **13**, 063002 (2018).
24. Baik, E. et al. Geospatial analysis of near-term potential for carbon-negative bioenergy in the United States. *Proc. Natl Acad. Sci. USA* **115**, 3290–3295 (2018).
25. Heck, V., Gerten, D., Lucht, W. & Popp, A. Biomass-based negative emissions difficult to reconcile with planetary boundaries. *Nat. Clim. Chang*. **8**, 151–155 (2018).
26. Amann, T. & Hartmann, J. Ideas and perspectives: synergies from co-deployment of negative emissions technologies. *Biogeosciences* **16**, 2949–2960 (2019).
27. Mayer, A. et al. The potential of agricultural land management to contribute to lower global surface temperature. *Sci. Adv*. **4**, eaaq0932 (2018).
28. Groffman, P. M. et al. Calcium additions and microbial nitrogen cycle processes in a northern hardwood forest. *Ecosystems* **9**, 1289–1305 (2006).
29. Dietzen, C., Harrison, R. & Michelsen-Correa, S. Effectiveness of enhanced mineral weathering as a carbon sequestration tool and alternative to agricultural lime: an incubation experiment. *Int. J. Greenhouse Gas Control* **74**, 251–258 (2018).
30. Smith, P., Haszeldine, R. S. & Smith, S. M. Preliminary assessment of the potential for, and limitations to, terrestrial negative emissions technologies in the UK. *Environ. Sci. Process. Impacts* **18**, 1400–1405 (2016).
31. DeLucia, E., Kantola, I., Blanc-Betes, E., Bernacchi, C. & Beerling, D. J. Basalt application for carbon sequestration reduces nitrous oxide fluxes from cropland. *Geophys. Res. Abstr*. **21**, EGU2019–EGU4500 (2019).
32. Das, S. et al. Cropping with slag to address soil, environment, and food security. *Front. Microbiol*. **10**, 1320 (2019).
33. Rogelj, J. et al. Paris Agreement climate proposals need a boost to keep warming well below 2 °C. *Nature* **534**, 631–639 (2016).
34. Clark, P. U. et al. Consequences of twenty-first-century policy for multi-millennial climate and sea-level change. *Nat. Clim. Chang*. **6**, 360–369 (2016).
35. Crowder, D. W. & Reganold, J. P. Financial competitiveness of organic agriculture on a global scale. *Proc. Natl Acad. Sci. USA* **112**, 7611–7616 (2015).
36. Bebbington, A. J. & Bury, J. T. Institutional challenges for mining and sustainability in Peru. *Proc. Natl Acad. Sci. USA* **106**, 17296–17301 (2009).
37. Renforth, P. et al. Silicate production and availability for mineral carbonation. *Environ. Sci. Technol*. **45**, 2035–2041 (2011).
38. Renforth, P. The negative emission potential of alkaline materials. *Nat. Commun*. **10**, 1401 (2019).
39. Tubana, B. S., Babu, T. & Datnoff, L. E. A review of silicon in soils and plants and its role in US agriculture: history and future perspectives. *Soil Sci*. **181**, 393–411 (2016).
40. Washbourne, C.-L. et al. Rapid removal of atmospheric CO_2 in urban soils. *Environ. Sci. Technol*. **49**, 5434–5440 (2015).
41. Lekakh, S. N. et al. Kinetics of aqueous leaching and carbonization of steelmaking slag. *Metallurg. Mater. Trans. B* **39**, 125–134 (2008).
42. Haynes, R. J., Belyaeva, O. N. & Kingston, G. Evaluation of industrial waste sources of fertilizer silicon using chemical extractions and plant uptake. *J. Plant Nutr. Soil Sci*. **176**, 238–248 (2013).
43. Rodd, A. V. et al. Surface application of cement kiln dust and lime to forage land: effect on forage yield, tissue concentration and accumulation of nutrients. *Can. J. Soil Sci*. **90**, 201–213 (2010).
44. Ramos, C.G. et al. Evaluation of soil re-mineralizer from by-product of volcanic rock mining: experimental proof using black oats and maize crops. *Nat. Res. Res*. 10.1007/s11053-019-09529-x (2019).
45. Savant, N. K., Datnoff, L. E. & Snyder, G. H. Depletion of plant-available silicon in soils: a possible cause of declining rice yields. *Commun. Soil Sci. Plant Anal*. **28**, 1245–1252 (1997).
46. Ning, D. et al. Impacts of steel-slag-based fertilizer on soil acidity and silicon availability and metals-immobilization in a paddy soil. *PLoS One* **11**, e0168163 (2016).
47. Chen, J. Rapid urbanization in China: a real challenge to soil protection and food security. *Catena* **69**, 1–15 (2007).
48. United Nations *Global Land Outlook* 1st edn (United Nations Convention to Combat Desertification, 2017).
49. Smith, M. R. & Myers, S. S. Impact of anthropogenic CO_2 emissions on global human nutrition. *Nat. Clim. Chang*. **8**, 834–839 (2018).
50. Cui, Z. et al. Pursuing sustainable productivity with millions of smallholder farmers. *Nature* **555**, 363–366 (2018).
51. Pidgeon, N. F. & Spence, E. Perceptions of enhanced weathering as a biological negative emissions option. *Biol. Lett*. **13**, 20170024 (2017).
52. Daval, D., Calvarusa, C., Guyut, F. & Turpault, M.-P. Time-dependent feldspar dissolution rates resulting from surface passivation: experimental evidence and geochemical implications. *Earth Planet. Sci. Lett*. **498**, 226–236 (2018).
53. Ricke, K., Drout, L., Caldeira, K. & Tavoni, M. Country-level social cost of carbon. *Nat. Clim. Chang*. **8**, 895–900 (2018).
54. Cox, E., Pidgeon, N. F., Spence, E. M. & Thomas, G. Blurred lines: the ethics and policy of greenhouse gas removal at scale. *Front. Environ. Sci*. https://doi.org/10.3389/fenvs.2018.00038 (2018).

Methods

CDR simulation framework

Our analysis is based on a one-dimensional vertical reactive transport model for rock weathering with steady-state flow[55,56], and a source term representing rock grain dissolution within the soil profile (Supplementary Methods). The model accounts for changing dissolution rates with soil depth and time as grains dissolve, and chemical inhibition of dissolution as pore fluids approach equilibrium with respect to the reacting basaltic mineral phases, and the formation of pedogenic calcium carbonate mineral in equilibrium with pore fluids. Simulations consider basalts exhibiting relatively slow- versus fast-dissolution rates due to differing mineralogy (Supplementary Tables 1–3). Basaltic minerals undergo dissolution at different rates, with some minerals continuing to undergo dissolution and to capture CO_2 after the first year of application. Thus calculating representative annual CDR rates requires computing average rates derived from repeated basaltic rock dust applications (Extended Data Fig. 7).

Transport equation. The calculated state variable in the transport equation is the dissolved molar equivalents of elements released by stoichiometric dissolution of mineral i, in units of moles per litre. ϕ is volumetric water content, C_i is dissolved concentration (moles per litre) of mineral i transferred to solution, t is time (years), q is vertical water flux (metres per year), z is distance along vertical flow path (metres), R_i is the weathering rate of basalt mineral i (moles per litre of bulk soil per year) and $C_{\mathrm{eq},i}$ is the solution concentration of weathering product at equilibrium with the mineral phase i (equation (1)):

$$\phi\frac{\partial C_i}{\partial t} = -q\frac{\partial C_i}{\partial z} + R_i\left(1-\frac{C_i}{C_{\mathrm{eq},i}}\right) \tag{1}$$

Mineral mass balance. The change in mass of basalt mineral i, B_i, is defined by the rate of stoichiometric mass transfer of the elements in mineral i to solution. Equation (2) is required because we are considering a finite mass of weathering rock, which over time can react to completion, as opposed to in situ weathering of the lithosphere, for example, when considering weathering and geomorphology[56].

$$\frac{\partial B_i}{\partial t} = -R_i\left(1-\frac{C_i}{C_{\mathrm{eq},i}}\right) \tag{2}$$

Removal of weathering products. The total mass balance over time for basalt mineral weathering allows calculation of the products transported from the soil profile. The total mass of weathering basalt is defined as follows, where m is the total number of weathering minerals in the rock, t_f is the duration of weathering (years) and L is the total depth of the soil profile (metres).

$$\text{Total mass of weathered basalt } = \sum_{i=1}^{m}\phi\int_{z=0}^{L}C_i(t,z)\,\mathrm{d}z + q\int_{t=0}^{t_f}C_i(t,L)\,\mathrm{d}t \tag{3}$$

We define q as the net annual sum of water gained through precipitation[57] and irrigation[58], minus crop evapotranspiration[59], as calculated with high-spatial-resolution gridded datasets of the contemporary climate in this initial analysis, given the uncertainties in infiltration and irrigation patterns at equivalent high spatial resolution for 2050 (Extended Data Figs. 8, 9; Supplementary Table 14).

Rate law. We modelled application of a crushed fast- or slow-weathering basalt, with specified mineral weight fractions and physico-chemical characteristics (Supplementary Tables 1–3). Rates of basalt grain weathering define the source term for weathering products and are calculated as a function of soil pH, soil temperature, soil hydrology and crop net primary productivity (NPP) using the linear transition state theory rate law[60–62]. Plant-enhanced basalt weathering is modelled empirically for annual and woody crops with power functions fitted to data (Supplementary Fig. 4; Supplementary Table 4). These functions represent the effects of a range of rhizosphere processes that accelerate the physical breakdown and chemical dissolution of minerals, including the activities of nutrient-scavenging mycorrhizal fungi that physically disrupt and chemically etch mineral surfaces, and bio-production of low-molecular-weight organic compounds and chelating agents[63,64].

Soil pH of each grid cell is dynamically calculated from the alkalinity mass and flux balance for an adaptive time-step, controlled by mineral dissolution rates, following initialization with a topsoil (0–15 cm) pH value based on field data from global soil databases (Supplementary Table 14); soil pH buffering capacity is accounted for with an empirical buffer function[65] relating soil pH to alkalinity. The soil p_{CO_2} depth profile of a grid cell is generated with the standard gas diffusion equation[66], scaled by crop NPP × 1.5 to account for combined autotrophic and heterotrophic respiration[67]. The alkalinity balance considers net acidity input during crop growth for biomass cations removed from the field[68], and secondary mineral precipitation of calcite[18].

Model advances

We incorporate three further relevant advances into the above one-dimensional vertical transport model with steady-state flow. First, we provide a numerical basis for calculating weathering rates using log-normal particle size distributions of basalt grains produced by mechanical crushing and grinding for soil amendment[22,69,70]. This conceptualization improves on the simplified case of a single mean particle diameter, previously used in ERW calculations[16–18,22,23]. Second, we apply the fractal dimension for surface roughness to relate reacting surface area to basalt mass across physical scales of weathering from the laboratory (in which the weathering kinetic parameter values are empirically determined) to the field (where model results reflect CDR operations)[71]. The fractal dimension effectively provides a means of consolidating measurements taken at different scales, and accounts for uncertainties in grain topography and porosity[72] that affect mass transfer rates from rock grains to flowing soil water. Finally, we calculate mean rates of rock dust weathering and CDR following annual applications by tracking cohorts of particles applied over a 10-year time horizon and their mineral composition (Extended Data Fig. 7).

Baseline simulations

Using this modelling framework, we analysed a baseline application rate of 40 tonnes per hectare per year (equivalent to a <2 mm layer of rock powder distributed over the croplands), which falls within the range of basalt amendments shown to improve crop production in field trials[4]. Net CDR is defined as the difference between CO_2 capture by ERW as dissolved inorganic carbon and soil (pedogenic) carbonate and the sum of CO_2 emissions for logistical operations. Carbon emissions per unit mass of ground rock depend on particle size (Extended Data Fig. 10), the CO_2 emissions per kilowatt hour of electricity generated from component energy sources (fossil fuels, nuclear and renewables), as well as the carbon costs of sourcing and transporting the silicate materials. Rock grinding to reduce particle size and maximize CDR is the primary energy-consuming operation in ERW[22,23,73].

Assessment of basalt transport from source regions to croplands is based on road and rail network analyses to calculate distances, costs and carbon emissions for each scenario (Supplementary Methods section 2.3). Our approach improves on prior analyses, which assumed a fixed radius between rock dust source and site of application[73]. We go beyond global cost estimates[23] by using national fuel (diesel), labour and infrastructure costs to undertake logistical operations, and the price of energy inputs to grind rocks. Our analysis thus represents the first techno-economic assessment in which detailed ERW carbon and economic costs vary within and between nations and account for socio-technical uncertainties in energy production.

CDR

We calculate CDR by ERW of crushed basalt applied to soils via two pathways: (1) the transfer of weathered base cations (Ca^{2+}, Mg^{2+}, Na^+ and K^+) from soil drainage waters to surface waters that are charge-balanced by the formation of HCO_3^- ions and transported to the ocean (equation (4)), and (2) formation of pedogenic carbonates (equation (5)).

Pathway 1 for calcium ions:

$$CaSiO_3 + 2CO_2 + 3H_2O \rightarrow Ca^{2+} + 2HCO_3^- + H_4SiO_4 \tag{4}$$

Pathway 2 for calcium carbonate formation:

$$Ca^{2+} + 2HCO_3^- \rightarrow CaCO_3 + CO_2 + H_2O \tag{5}$$

Monovalent and divalent base cations are released from basaltic minerals by dissolution based on stoichiometry (Supplementary Table 2). CDR, via pathway 1, potentially sequesters two moles of CO_2 from the atmosphere per mole of divalent cation. However, ocean carbonate chemistry reduces the efficiency of CDR (η) to an extent depending on ocean temperature, salinity and the surface ocean dissolved CO_2 concentration. We calculate η for average ocean temperature (17 °C), salinity (35%) and a Representative Concentration Pathway (RCP) 8.5 simulation for 2050 (the worst-case scenario) for dissolved p_{CO_2} of 600 µatm, giving $\eta = 0.86$, that is, 0.86 moles of CDR per mole of monovalent cation and 1.72 moles of CO_2 removed per mole of divalent cation added to the oceans[20]. For pathway 1, the efficiency of CDR = $\eta \times \sum$(moles of monovalent cations) + $2\eta \times \sum$(moles of divalent cations).

CDR via pathway 2 can occur if dissolved inorganic carbon derived from atmospheric CO_2 precipitates as pedogenic carbonate, and sequesters 1 mole of CO_2 per mole of Ca^{2+} instead of 1.72 moles of CO_2 via pathway 1 and is therefore less efficient. Thus for any given grid cell, we compute CDR by ERW as the alkalinity flux in soil drainage and pedogenic calcite precipitation. Possible CO_2 degassing due to changes in surface water chemistry during transport in large river systems[74] is not considered.

Cost assessment modelling

An overview of the environmental costs model and its linkages with the performance model is presented in Extended Data Fig. 4. We include contributions to total cost of (1) mining, (2) processing[75,76], (3) distribution and transport and (4) spreading on agricultural land. We considered how the cost of energy and the carbon emissions varied with grinding to different particle size distributions (Extended Data Fig. 10). Grinding to finer particles requires greater energy and results in higher carbon emissions. We defined the particle size distribution by the p80 value; p80 is defined as 80% of the particles having a diameter less than or equal to a specified particle size. We calculated the optimized p80 that results in maximum net CDR for each grid cell and this was conducted for different fractions of a country's crop area (0.1 to 1.0 at 0.1 increments), ordered according to weathering potential. For a given p80 value, we calculate the weathering rate for each grid cell, sort them in descending order and find the grid cells that comprise the cumulative fraction of total land for each incremental increase in land area.

Optimization is conducted by country for each of the two types of basalt and their log-normal particle size distributions (Supplementary Tables 1–3). Country-specific electricity production and the forecast fractional contributions to electricity production by different energy sources (coal, natural gas, oil, solar photovoltaics, concentrated solar power, hydropower, wind, marine) for 2050 are based on BAU, that is, currently implemented energy policies, and energy projections consistent with a 2 °C warming scenario (Extended Data Fig. 9)[5]. National CO_2 emissions for electricity generation consistent with both scenarios were based on results reported in ref. [5] (Supplementary Tables 6–9). Industrialized nations (for example, Canada) consume up to about 2% of their total energy production on rock comminution (crushing and grinding) processes[77]. We assume a future maximum upper limit of 3% energy consumption for all nations, based on the rationale that current rates for developed nations grow from around 2% today in line with national projected energy production[5] in 2050 (Extended Data Fig. 2).

Distribution costs and emissions were calculated by performing spatial analysis with ArcGIS (http://www.arcgis.com) software. Basalt rock sources were identified from the GLiM rock database[78], excluding those in protected areas[79]. We then performed a global transport (rail and road) network analysis by modelling a logistic ERW supply by creating an origin–destination cost matrix using GIS[80,81]. For larger datasets, the origin–destination cost matrix searches and measures the least-cost paths along the network from multiple origins to multiple destinations to identify the most cost-effective or shortest route between a source and destination. Transport analyses used the lowest-emission option between rail and road network to calculate distribution costs and CO_2 emissions (Supplementary Tables 10–12). Freight-rail emissions were obtained from 2050 projections of reduced carbon emissions following improvements in energy efficiency[82]. Rail CO_2 emissions were the same for both the BAU and 2 °C scenarios. For road transport, we considered estimated energy consumption of currently or soon-to-be available heavy electric trucks 1.38 (kWh km^{-1})[83] and projected carbon emissions in the electricity sector of each country for BAU or the 2 °C scenario[5].

Forecasting bulk silicate waste production

We developed a model that relates global per capita material production (for cement) or consumption (steel) P to per capita gross world product (GWP)[84,85] through historical global data using nonlinear least squares (equation (6)):

$$P = a\mathrm{e}^{-b/\mathrm{GWP}} \tag{6}$$

where a and b are regression constants. The derived saturation value, a, was used in a further regression through national data normalized to 2014 production and GDP (equation (7).

$$P = P_{\mathrm{ref}} \times [1 + (m + r) \times \Delta\mathrm{GDP}] \times \mathrm{e}^{a \times (1 - \mathrm{e}^{-m \times \Delta\mathrm{GDP}}) - (m \times \Delta\mathrm{GDP})} \tag{7}$$

where P_{ref} is the global per capita consumption in a given reference year (2014), ΔGDP is the deviation of the per capita gross domestic product from the reference year, and m and r are regression constants. These results were used together with averaged projections of future GDP (Supplementary Table 14) from the 'middle-of-the-road' Shared Socioeconomic Pathway (SSP2) to derive nationally resolved projections of future per capita consumption/production[85]. SSP2 potentially represents the largest material production pathway, as other SSPs forecast lower consumption or economic growth producing 30% to 50% less material globally. We have not considered the penetration of recycling into steel production beyond its current rate. Cement and cement kiln dust have no capacity to be recycled as cement. The total production/consumption at a given time, $T(t)$, was calculated by multiplying the population, $\mathrm{Pop}(t)$, by production or consumption (P). We assume that 115 kg of cement kiln dust is produced as a by-product in kilns for every tonne of clinker, and have modelled the production of demolition waste following an average 50-year service life (normally distributed with a standard deviation of 10 years)[86]. The ratio of pig iron to steel production (0.72) was obtained using linear regression of 1960–2014 data, negating the need to explicitly model pig iron displacement from scrap recycling, and assuming that the scrap ratio remains unchanged. All steel and blast furnace slag was considered available for reaction with CO_2. Between 2006 and 2014, 185 kg of blast furnace slag and 117 kg of steel slag was produced for every tonne of crude steel[87].

Data availability

Datasets on global crop production and yield are available at http://www.earthstat.org/, accessed on 18 December 2019. Datasets on global crop irrigation are available at https://zenodo.org/record/1209296, accessed on 18 December 2019. Datasets on global precipitation are available at http://www.climatologylab.org/terraclimate.html, accessed on 18 December 2019. Datasets on global soil surface pH are available at https://daac.ornl.gov/SOILS/guides/HWSD.html, accessed on 18 December /2019. Datasets on global soil temperature are available at https://esgf-node.llnl.gov/search/cmip5/, accessed on 18 December 2019. Datasets on diesel prices are available at https://data.worldbank.org/indicator/EP.PMP.DESL.CD. Datasets on mining costs are available at http://www.infomine.com/. Datasets on gross national income per capita are available at https://data.worldbank.org/indicator/ny.gnp.pcap.pp.cd. Datasets for projections of future GDP linked to Shared Socioeconomic Pathways are available at https://tntcat.iiasa.ac.at/SspDb. Source data are provided with this paper.

Code availability

The Matlab codes developed for this study belong to the Leverhulme Centre for Climate Change Mitigation. The authors will make them available upon reasonable request.

55. Berner, R. A. Rate control of mineral dissolution under Earth surface conditions. *Am. J. Sci*. **278**, 1235–1252 (1978).
56. Maher, K. The dependence of chemical weathering rates on fluid residence time. *Earth Planet. Sci. Lett*. **294**, 101–110 (2010).
57. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci. Data* **5**, 170191 (2018).
58. Huang, Z. W. et al. Reconstruction of global gridded monthly sectoral water withdrawals for 1971-2010 and analysis of their spatiotemporal patterns. *Hydrol. Earth Syst. Sci*. **22**, 2117–2133 (2018).
59. Siebert, S. & Doll, P. Quantifying blue and green virtual water contents in global crop production as well as potential production losses without irrigation. *J. Hydrol*. **384**, 198–217 (2010).
60. Aagaard, P. & Helgeson, H. C. Thermodynamic and kinetic constraints on reaction-rates among minerals and aqueous-solutions. 1. *Theoretical considerations. Am. J. Sci*. **282**, 237–285 (1982).
61. Lasaga, A. C. Chemical-kinetics of water-rock interactions. *J. Geophys. Res*. **89**, 4009–4025 (1984).
62. Brantley, S. L., Kubicki, J. D. & White, A. F. *Kinetics of Water–Rock Interaction* (Springer, 2008).
63. Harley, A. D. & Gilkes, R. J. Factors influencing the release of plant nutrient elements from silicate rock powders: a geochemical overview. *Nutr. Cycl. Agroecosyst*. **56**, 11–36 (2000).
64. Taylor, L. L. et al. Biological evolution and the long-term carbon cycle: integrating mycorrhizal evolution and function into the current paradigm. *Geobiology* **7**, 171–191 (2009).
65. Nelson, P. N. & Su, N. Soil pH buffering capacity: a descriptive function and its application to some acidic tropical soils. *Aust. J. Soil Sci*. **48**, 201–207 (2010).
66. Cerling, T. Carbon dioxide in the atmosphere: evidence from Cenozoic and Mesozoic paleosols. *Am. J. Sci*. **291**, 377–400 (1991).
67. Taylor, L., Banwart, S. A., Leake, J. R. & Beerling, D. J. Modelling the evolutionary rise of ectomycorrhizal on sub-surface weathering environments and the geochemical carbon cycle. *Am. J. Sci*. **311**, 369–403 (2011).
68. Banwart, S. A., Berg, A. & Beerling, D. J. Process-based modeling of silicate mineral weathering responses to increasing atmospheric CO_2 and climate change. *Glob. Biogeochem. Cycles* **23**, GB4013 (2009).
69. Petavratzi, E., Kingman, S. & Lowndes, I. Particulates from mining operations: a review of sources, effects and regulations. *Miner. Eng*. **18**, 1183–1199 (2005).
70. Cepuritis, R., Garboczi, E. J., Ferraris, C. F., Jacobsen, S. & Sorensen, B. E. Measurement of particle size distribution and specific surface area for crushed concrete aggregate fines. *Adv. Powder Technol*. **28**, 706–720 (2017).
71. Navarre-Sitchler, A. & Brantley, S. Basalt weathering across scales. *Earth Planet. Sci. Lett*. **261**, 321–334 (2007).
72. Brantley, S. L. & Mellott, N. P. Surface area and porosity of primary silicate minerals. *Am. Mineral*. **85**, 1767–1783 (2000).
73. Moosdorf, N., Renforth, P. & Hartmann, J. Carbon dioxide efficiency of terrestrial weathering. *Environ. Sci. Technol*. **48**, 4809–4816 (2014).
74. Salisbury, J. E. et al. Seasonal observations of surface waters in two Gulf of Maine estuary-plume systems: relationships between watershed attributes, optical measurements and surface p_{CO2}. *Estuar. Coast. Shelf Sci*. **77**, 245–252 (2008).
75. Darling, P. & Society for Mining, Metallurgy and Exploration (U.S.). *SME Mining Engineering Handbook* 3rd edn (Society for Mining, Metallurgy and Exploration, 2011).
76. *InfoMine, Mining Cost Service* http://www.infomine.com/ (Infomine, 2009).
77. Tromans, D. Mineral comminution: energy efficiency considerations. *Min. Eng*. **21**, 613–620 (2008).
78. Hartmann, J. & Moosdorf, N. The new global lithological map database GLiM: a representation of rock properties at the Earth surface. *Geochem. Geophys. Geosyst*. **13**, Q12004 (2012).
79. Protected Planet: The World Database on Protected Areas (WDPA)/The Global Database on Protected Areas Management Effectiveness (GD-PAME) https://www.protectedplanet.net/ (UNEP-WCMC and IUCN, 2018).
80. ROTARU, A. S. et al. Modelling a logistic problem by creating an origin-destination cost matrix using GIS technology. *Bull. UASVM Horticulture* **71**, https://doi.org/10.15835/buasvmcn-hort:9697 (2014).
81. Osorio, C. Dynamic origin-destination matrix calibration for large-scale network simulators. *Transport. Res. C* **98**, 186–206 (2019).
82. *International Energy Agency The Future of Rail, Opportunities for Energy and the Environment* (International Energy Agency, 2019).
83. Liimatainen, H., van Vliet, O. & Aplyn, D. The potential of electric trucks—an international commodity-level analysis. *Appl. Energy* **236**, 804–814 (2019).
84. *GDP (current US$)* https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (The World Bank, 2016).
85. Bauer, N. et al. Shared socio-economic pathways of the energy sector – quantifying the narratives. *Glob. Environ. Change* **42**, 316–330 (2017).
86. Xi, F. et al. Substantial global carbon uptake by cement carbonation. *Nat. Geosci*. **9**, 880–883 (2016).
87. U.S. Geological Survey. *Mineral Commodity Summaries 2006* (US Geological Survey, 2006).

Acknowledgements We thank A. Azapagic and J. Shepherd for comments on an earlier draft, and acknowledge discussions with additional members of the Royal Society-Royal Academy of Engineering Greenhouse Gas Removal Working Group. We acknowledge funding of this research with a Leverhulme Research Centre Award (RC-2015-029) from the Leverhulme Trust. We thank L. Taylor for advice and discussions during model development and J. Quirk for data and analysis on plant weathering. P.R. acknowledges UKRI funding under the UK Greenhouse Gas Removal Programme (NE/P019943/1, NE/P019730/1); I.A.J. acknowledges financial support from the Research Council of the University of Antwerp. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling responsible for CMIP and thank the climate modelling groups for producing and making available their model output. For CMIP, the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Author contributions D.J.B., E.P.K., M.R.L., P.W., S.Q. and S.A.B. designed the study, E.P.K. and M.R.L. undertook model development and coding, with input from P.W., S.A.B., S.Q. and D.J.B. E.P.K. undertook data analysis and synthesis, R.M.E. and L.K. undertook the GIS transport analyses, P.R. did the silicate production modelling, and N.F.P. wrote sections on public perception. J.-F.M., H.P., N.R.E. and P.B.H. provided data on national energy production and sources, and CO_2 emissions for both scenarios. M.G.A., R.H.J., C.R.P., M.K., B.S. and I.A.J. all provided input on sections and addition of appropriate references specific to their area of expertise. D.J.B. and S.A.B. wrote the manuscript, with input from J.H.

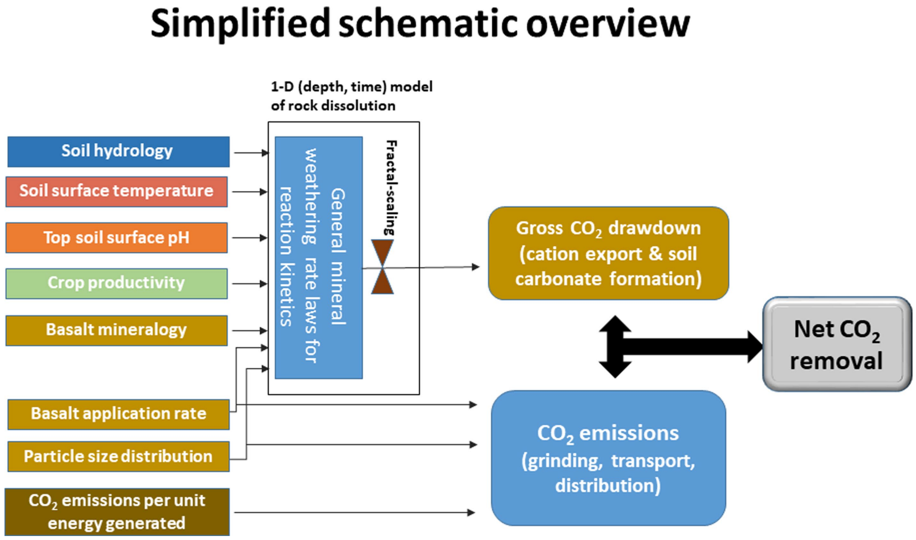
Competing interests The authors declare no competing interests.

Additional information

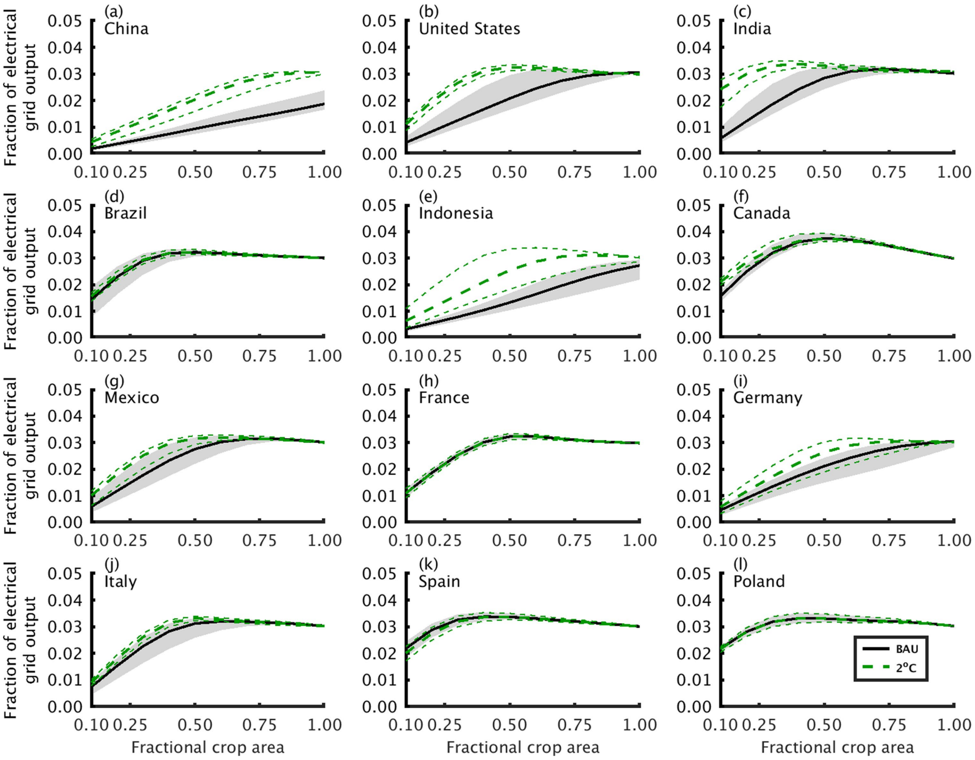
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2448-9.

Correspondence and requests for materials should be addressed to D.J.B.

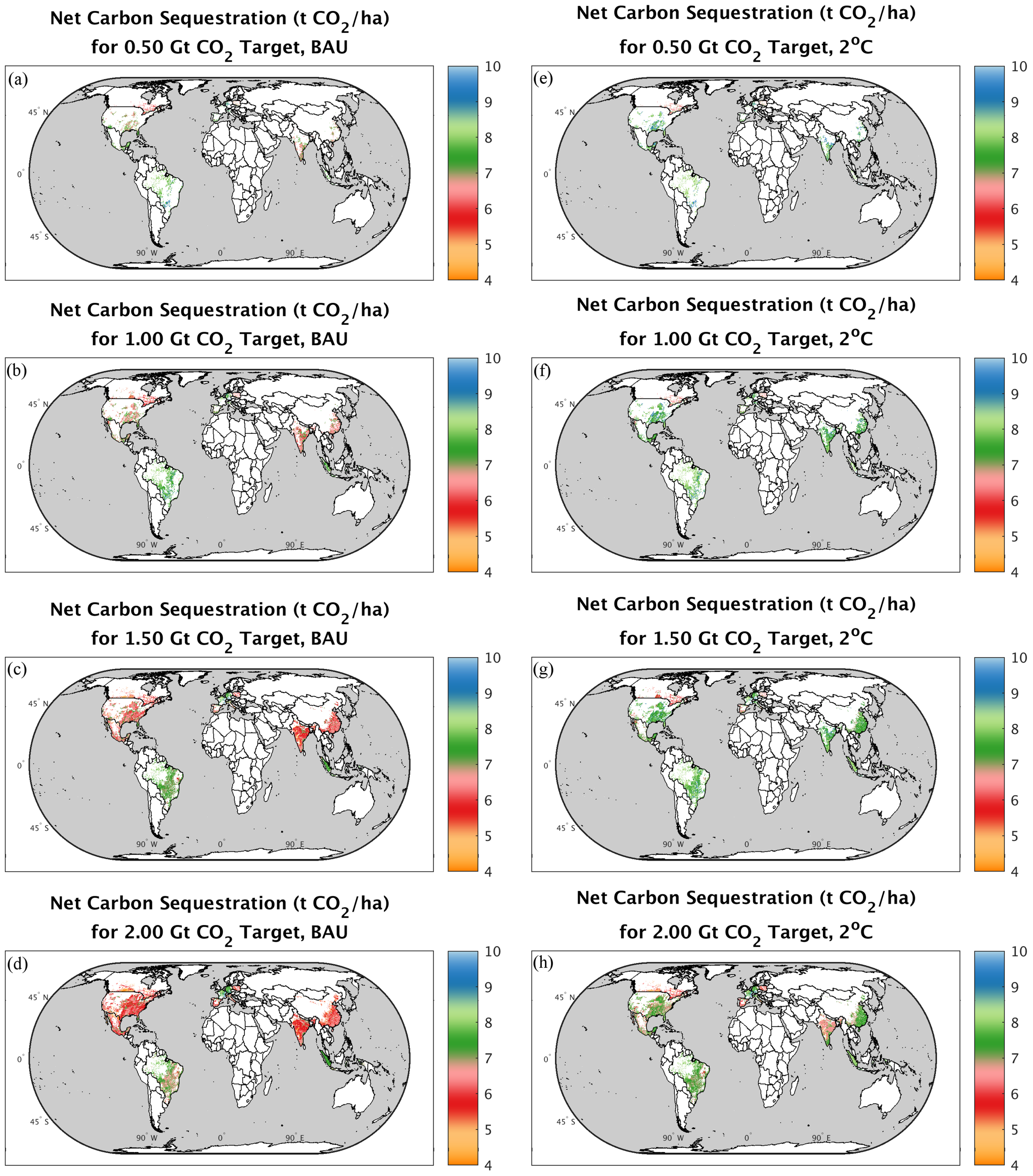
Peer review information *Nature* thanks Johannes Lehmann, Keith Paustian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Performance model schematic. Detailed methods are provided in Methods sections 'CDR simulation framework' and 'Model advances'. Spatially resolved key drivers are mapped in Extended Data Fig. 8; sources given in Supplementary Table 14.



Extended Data Fig. 2 | Cumulative energy demand for rock grinding by nation. Results are shown for the top seven nations of the world (**a**–**g**), and the top five European nations (**h**–**l**), as ranked by net CDR capacity, with increasing fractional cropland area of ERW deployment. Curves depict simulations for the BAU and 2 °C energy policy scenarios. The grey-shaded area for each nation represents the 90% confidence interval calculated for basalts with relatively slow- versus fast-weathering rates for the BAU scenario; short green dashed lines indicate the 90% confidence limits of the corresponding 2 °C scenario simulations.



Net Carbon Sequestration (t CO_2/ha) for 0.50 Gt CO_2 Target, BAU

Net Carbon Sequestration (t CO_2/ha) for 0.50 Gt CO_2 Target, 2°C

Net Carbon Sequestration (t CO_2/ha) for 1.00 Gt CO_2 Target, BAU

Net Carbon Sequestration (t CO_2/ha) for 1.00 Gt CO_2 Target, 2°C

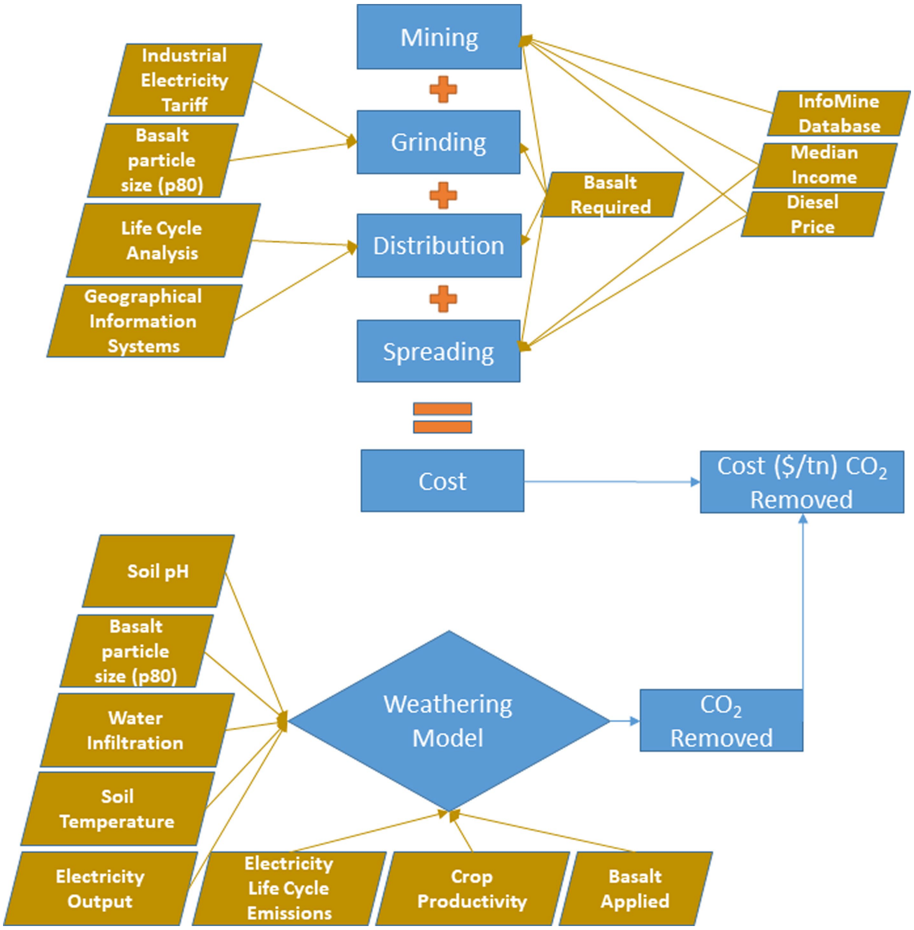
Net Carbon Sequestration (t CO_2/ha) for 1.50 Gt CO_2 Target, BAU

Net Carbon Sequestration (t CO_2/ha) for 1.50 Gt CO_2 Target, 2°C

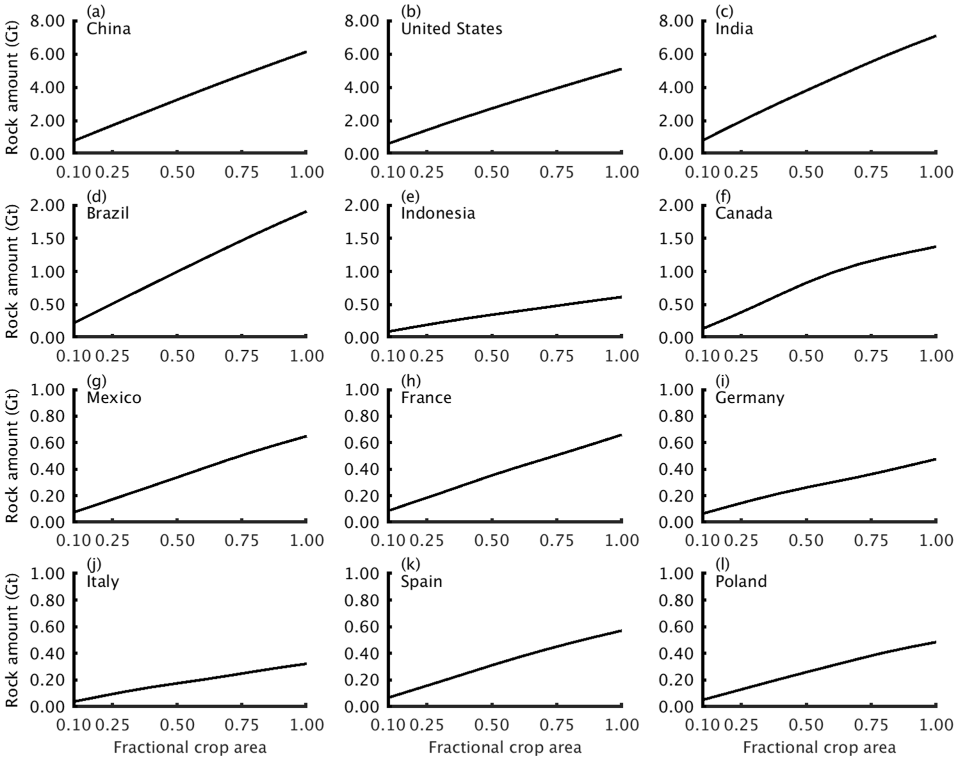
Net Carbon Sequestration (t CO_2/ha) for 2.00 Gt CO_2 Target, BAU

Net Carbon Sequestration (t CO_2/ha) for 2.00 Gt CO_2 Target, 2°C

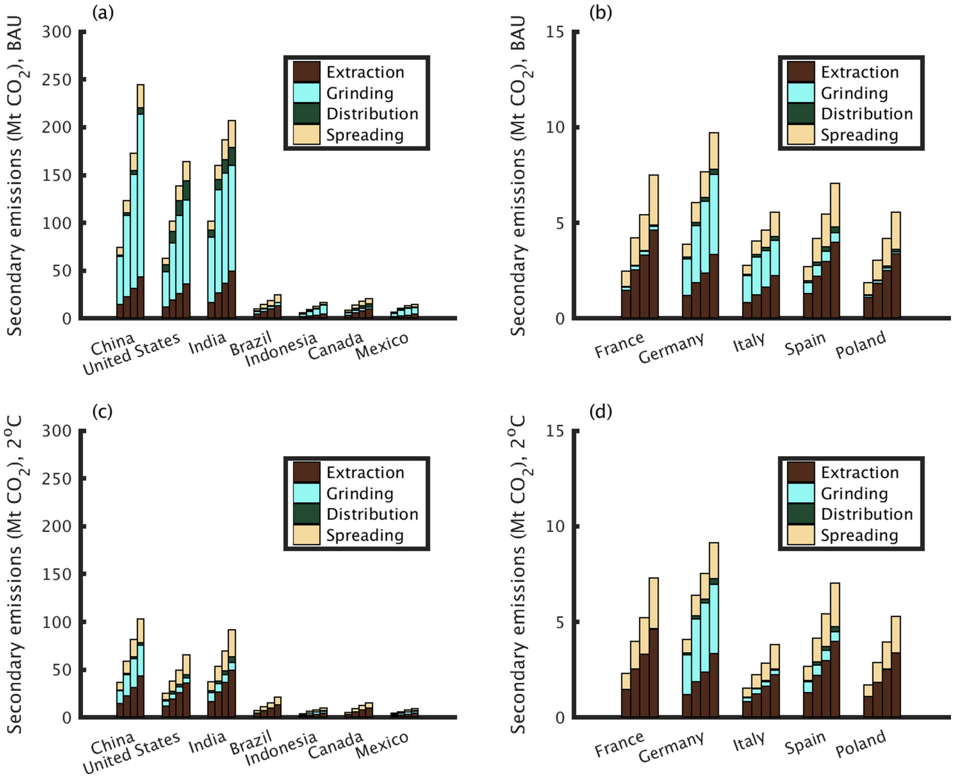
Extended Data Fig. 3 | Simulated net CDR with croplands via ERW. Net rates of CO_2 sequestration on croplands (annual and perennial combined) for the four targeted global CDR rates, 0.5 Gt CO_2 yr^{-1}, 1.0 Gt CO_2 yr^{-1}, 1.5 Gt CO_2 yr^{-1} and 2.0 Gt CO_2 yr^{-1} (Table 1) for the BAU (**a**–**d**) and the 2 °C (**e**–**h**) energy policy scenarios.



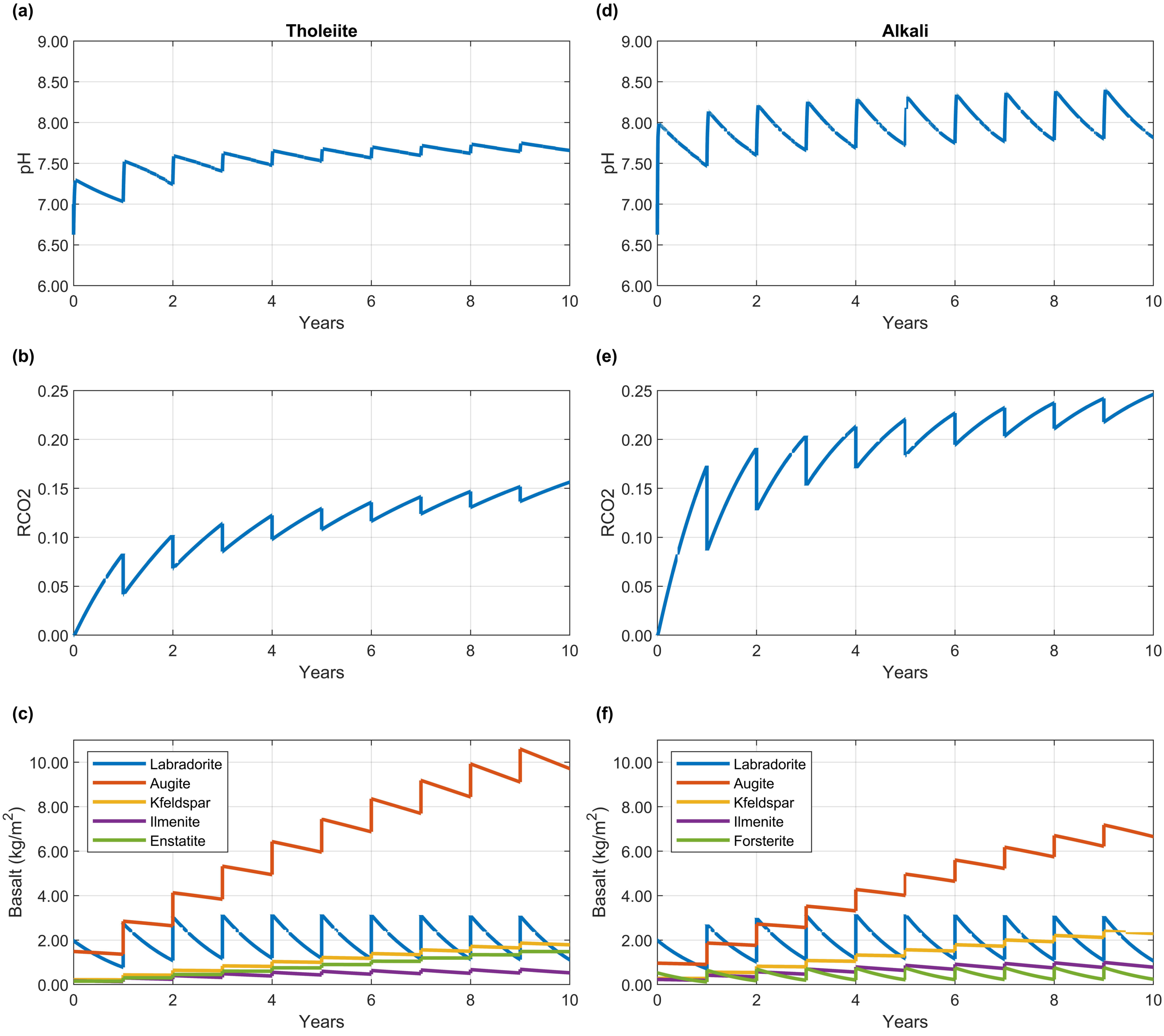
Extended Data Fig. 4 | Schematic overview of the environmental economics model. Interactions between the performance model, calculating net CDR, and the major components of the environmental economic model. Spatially resolved key drivers are mapped in Extended Data Fig. 9; sources given in Supplementary Table 14. Brown shading denotes inputs; blue shading denotes processes.



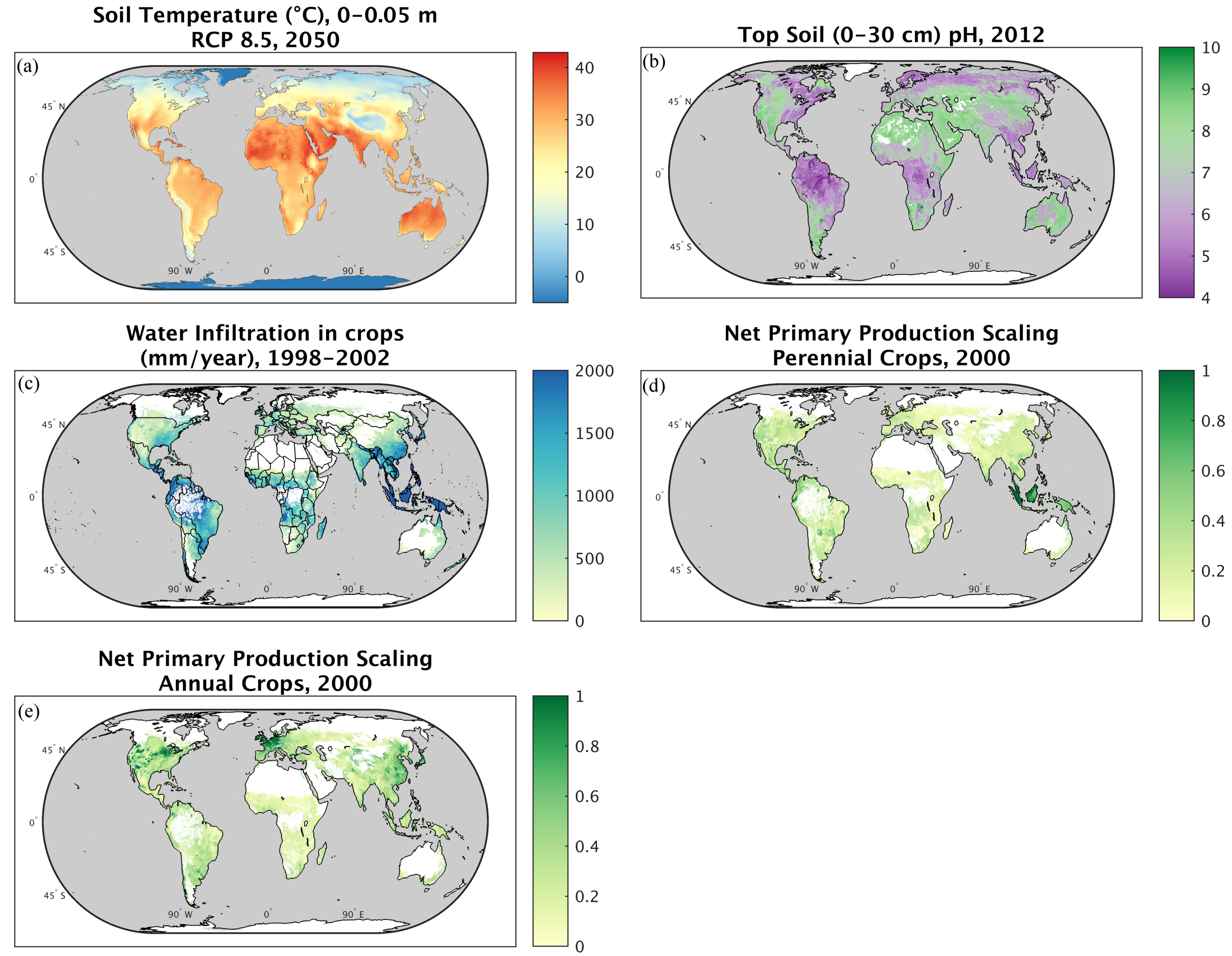
Extended Data Fig. 5 | Cumulative silicate demand by nation. Results are shown for the seven nations of the world (**a**–**g**) and the five European nations (**h**–**l**) with the highest CDR, as ranked by net CDR capacity, with increasing fractional cropland area deployment of ERW. Note the *y*-axis scale changes for European countries. Curves are the same irrespective of energy policy scenario.



Extended Data Fig. 6 | Secondary CO$_2$ emissions from logistical ERW operations in 2050. a–**d**, Results are shown for the seven nations of the world (**a**, **c**) and the five European nations (**b**, **d**) with the highest CDR potential for the BAU scenario (**a**, **b**) and for the 2 °C energy policy scenario (**c**, **d**). For each country, from left to right, bars are for fractions of 0.25, 0.5, 0.75 and 1.0 of ERW deployment on croplands. Under the BAU scenario, CO$_2$ emissions from grinding dominate secondary emissions associated with ERW, except for France, where low-carbon nuclear power dominates. Under the 2 °C energy policy scenario (**c** and **d**), secondary CO$_2$ emissions generally drop for most nations as they transition to low-carbon energy sources in 2050 and implement negative emissions.



Extended Data Fig. 7 | Multi-year performance model simulations of weathering. Illustrative multi-year simulations of annual basalt application with the performance model showing the effects on soil pH, average efficiency of CDR (RCO$_2$), and soil mineral masses over a 10-year time horizon. **a**–**c**, pH, RCO$_2$ and mineral mass results for the tholeiitic basalt, respectively. **d**–**f**, pH, RCO$_2$ and mineral mass results for the alkali basalt (Supplementary Tables 1–3). All simulations used the same p80 particle size (100 μm) and were undertaken at 20 °C. Multi-year simulations capture the effect of basaltic minerals undergoing dissolution at different rates, with some minerals continuing to undergo dissolution and capture CO$_2$ after the first year of application. Such simulations allow average rates of weathering and CDR from repeated basaltic rock dust applications to be computed. Our extended theory underpinning the simulation framework tracks cohorts of particles applied each year and their mineral composition over time to account for cumulative effects (Supplementary Methods).


Soil Temperature (°C), 0–0.05 m RCP 8.5, 2050

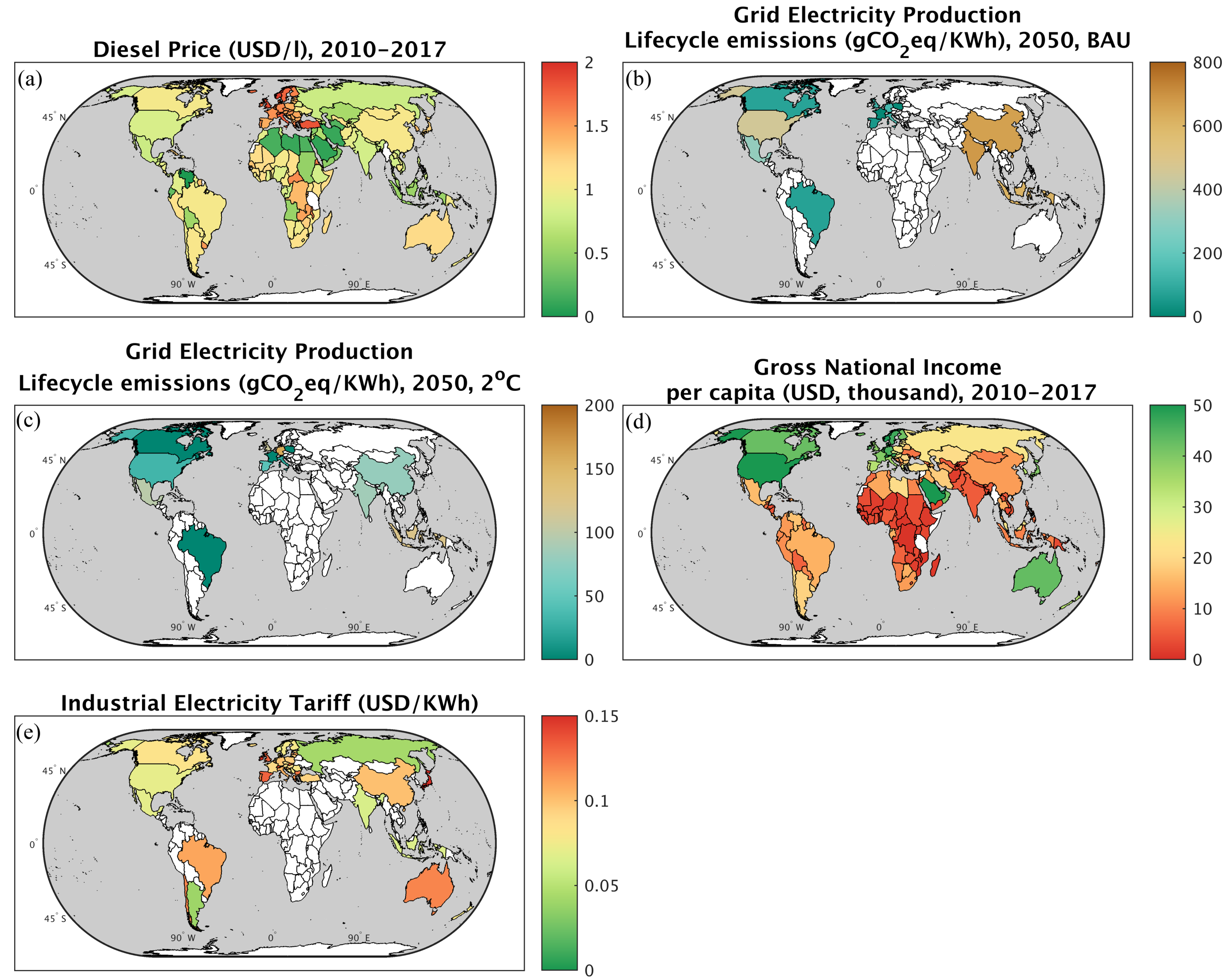
Top Soil (0–30 cm) pH, 2012

Water Infiltration in crops (mm/year), 1998–2002

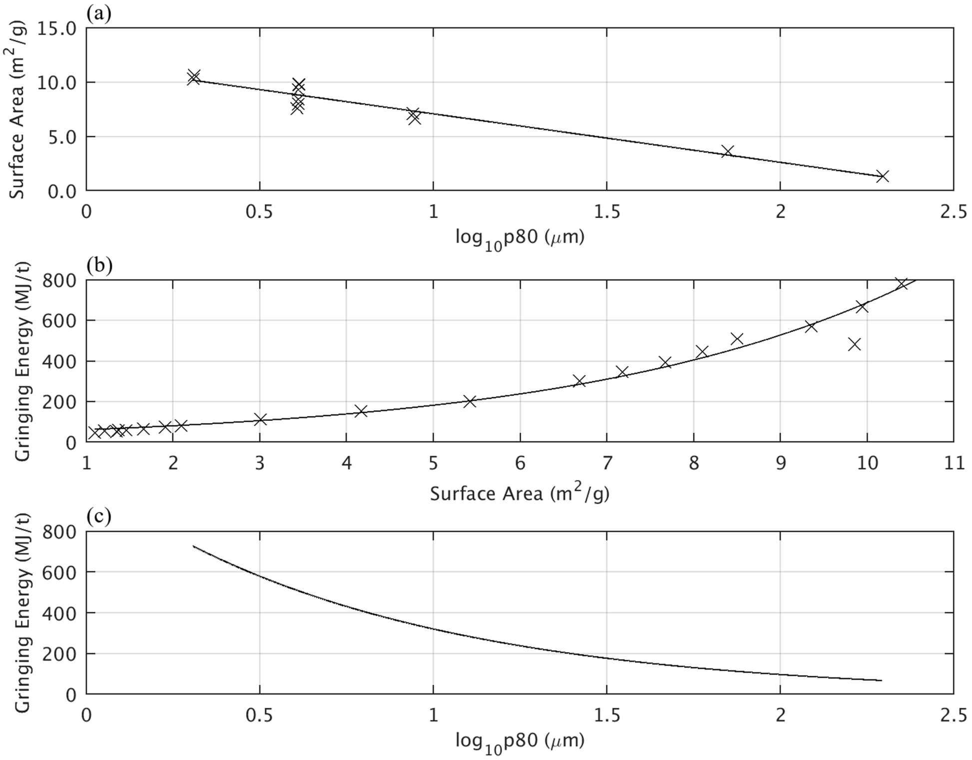
Net Primary Production Scaling Perennial Crops, 2000

Net Primary Production Scaling Annual Crops, 2000

Extended Data Fig. 8 | Spatially resolved drivers of the performance model. **a**, Soil temperature from the Hadley Centre coupled Earth System Model (HadGEM) RCP 8.5 simulation for 2050 (the worst-case scenario). **b**, The HYDE harmonized soil pH database. **c**, Annual cropland soil water infiltration (irrigation water + precipitation minus evapotranspiration). **d**, **e**, Net primary production index for perennial and annual crops as derived from FAO datasets, respectively. Data sources and spatial resolution are specified in Supplementary Table 14.



Extended Data Fig. 9 | Spatially resolved drivers for environmental economics modelling. **a**, Industrial diesel prices. **b**, **c**, CO_2 emissions intensity for the BAU scenario (**b**) and the 2 °C scenario (**c**). **d**, Gross national income per capita. **e**, Industrial electricity prices. Data sources and spatial resolution are specified in Supplementary Table 14.



Extended Data Fig. 10 | Relationship between particle size, surface area and grinding energy. a, Relationship between particle size and surface area. **b**, Relationship between surface area and grinding energy. **c**, Relationship between particle size and grinding energy. p80 is defined as 80% of the particles having a diameter less than or equal to the specified size. Derived from data in ref. [73].

A Cambrian crown annelid reconciles phylogenomics and the fossil record

https://doi.org/10.1038/s41586-020-2384-8

Received: 25 October 2019

Accepted: 26 March 2020

Hong Chen[1,2,3,9], **Luke A. Parry**[4,5,9], **Jakob Vinther**[6,7], **Dayou Zhai**[1,2], **Xianguang Hou**[1,2] ✉ & **Xiaoya Ma**[1,2,8] ✉

The phylum of annelids is one of the most disparate animal phyla and encompasses ambush predators, suspension feeders and terrestrial earthworms[1]. The early evolution of annelids remains obscure or controversial[2,3], partly owing to discordance between molecular phylogenies and fossils[2,4]. Annelid fossils from the Cambrian period have morphologies that indicate epibenthic lifestyles, whereas phylogenomics recovers sessile, infaunal and tubicolous taxa as an early diverging grade[5]. Magelonidae and Oweniidae (Palaeoannelida[1]) are the sister group of all other annelids but contrast with Cambrian taxa in both lifestyle and gross morphology[2,6]. Here we describe a new fossil polychaete (bristle worm) from the early Cambrian Canglangpu formation[7] that we name *Dannychaeta tucolus*, which is preserved within delicate, dwelling tubes that were originally organic. The head has a well-defined spade-shaped prostomium with elongated ventrolateral palps. The body has a wide, stout thorax and elongated abdomen with biramous parapodia with parapodial lamellae. This character combination is shared with extant Magelonidae, and phylogenetic analyses recover *Dannychaeta* within Palaeoannelida. To our knowledge, *Dannychaeta* is the oldest polychaete that unambiguously belongs to crown annelids, providing a constraint on the tempo of annelid evolution and revealing unrecognized ecological and morphological diversity in ancient annelids.

Annelida Lamarck, 1809
Palaeoannelida Weigert & Bleidorn, 2016
Magelonidae Cunningham & Ramage, 1888
Dannychaeta tucolus gen. et sp. nov.

Etymology. *Danny* refers to Danny Eibye-Jacobsen, for his contributions to our understanding of early annelids; *chaeta* (Latin), bristle; *tubus* (Latin), tube; *colus* (Latin), dwelling in.
Holotype. YKLP 11382 part and counterpart (YKLP, Yunnan Key Laboratory for Palaeobiology) (Fig. 1 and Extended Data Fig. 1).
Referred material. YKLP 11383–11402 (Figs. 2, 3 and Extended Data Figs. 2–6).
Horizon and locality. Canglangpu Formation, Cambrian stage 3, Hongjingshao Member (around 514 million years ago), at southwest of Guanshan reservoir, Chenggong, Kunming, China[7].
Diagnosis. Elongated, slender polychaetes with organic dwelling tubes. Head with anteriorly tapering spade-shaped prostomium, with paired palps attached ventrolaterally near the mouth. Body heteronomously segmented, with a wider thorax containing at least eight chaetigers. Parapodia with lateral lamellae in the posterior part of the abdomen. Abdominal parapodia biramous, unknown in thorax. Capillary chaetae in both rami, occurring in tight parallel bundles.

Description. The holotype (Fig. 1 and Extended Data Fig. 1) is incomplete posteriorly (around 40 mm long), has a wider anterior region (thorax; maximum width, 3.9 mm) (Fig. 1a–e) and abdomen (about 1.9 mm, excluding parapodia) (Fig. 1h, i). The number of thoracic segments would consist of at least 8 chaetigers, extrapolating from segment spacing (15.5 mm length and 1 segment per 1.9 mm). The prostomium is a spade-shaped lobe (Figs. 1c–e, 2 and Extended Data Figs. 1a–g, 2a–h) and is longer (approximately 4 mm) than it is wide (around 2 mm). The relief of overlapping anatomical features preserved on different planes indicates that the prostomium is dorsal of the palps (Figs. 1d–f, 2b–d). The palps cross over each other in the holotype (Fig. 1d–f), are incompletely preserved anteriorly, but are at least 30% the length of the thorax. A specimen in ventral view shows palps that insert ventrolaterally, anterior of a putative burrowing organ (Fig. 2a–d). The prostomium has a faint pair of tapering ridges (Fig. 2d). The gut is preserved as a carbonaceous film (Fig. 3f) that terminates adjacent to palp bases (Fig. 1d–f), indicating palp attachment near the mouth opening.

Abdominal parapodia are distinct lobes, projecting around 300 μm from the body (Fig. 3c–e). Anterior abdominal chaetae in the holotype are around 500 μm long (Fig. 1h). In narrower midbody fragments, chaetae are approximately 800 μm long (Fig. 3d), suggesting that the chaetation is variable along the body. Abdominal chaetae are in tight fascicles (Figs. 1j, 3c–e and Extended Data Fig. 3k) and are the clearest

---

[1] Yunnan Key Laboratory for Palaeobiology, Institute of Palaeontology, Yunnan University, Kunming, China. [2] MEC International Joint Laboratory for Palaeobiology and Palaeoenvironment, Institute of Palaeontology, Yunnan University, Kunming, China. [3] School of Biological Sciences and Technology, Liupanshui Normal University, Liupanshui, China. [4] Department of Geology and Geophysics, Yale University, New Haven, CT, USA. [5] Department of Earth Sciences, University of Oxford, Oxford, UK. [6] School of Earth Sciences, University of Bristol, Bristol, UK. [7] School of Biological Sciences, University of Bristol, Bristol, UK. [8] Centre for Ecology and Conservation, University of Exeter, Penryn, UK. [9] These authors contributed equally: Hong Chen, Luke A. Parry. ✉ e-mail: xghou@ynu.edu.cn; x.ma2@exeter.ac.uk

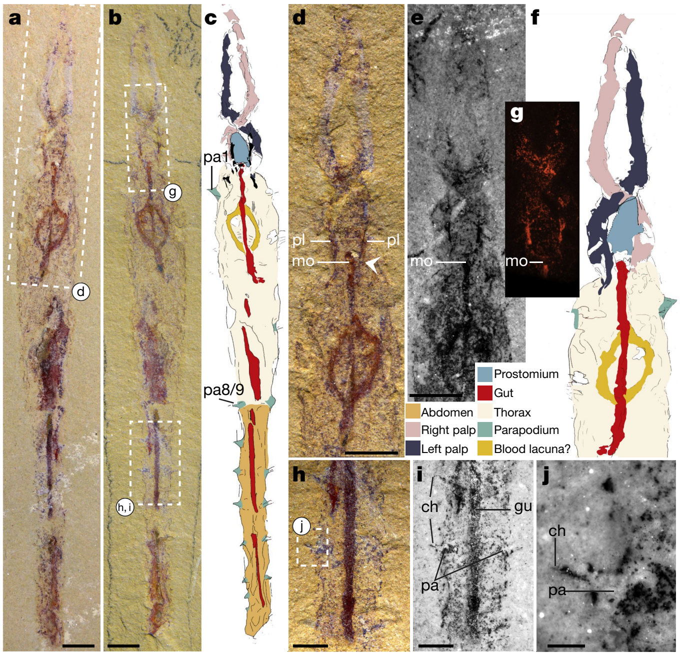


Fig. 1 | Holotype specimen YKLP 11382 of *D. tucolus*. **a**, **b**, Part (**a**) and counterpart (**b**), imaged using direct light. **c**, Interpretative drawing of the counterpart. The colour scheme is described in **f**. **d**, Anterior region of the part. **e**, Anterior region of the counterpart under fluorescent light. The specimen is mirrored to facilitate the comparison with the part. **f**, Interpretative drawing of the anterior, based on both part and counterpart. **g**, Fe map from SEM–energy-dispersive X-ray spectroscopy (SEM–EDX) analysis. **h**, **i**, Abdominal region, counterpart, obtained using direct light (**h**) and fluorescent light (**i**). **j**, Magnification of the parapodium and chaetae. ch, chaetae; gu, gut; mo, mouth; pa, parapodium; pa1, parapodium of segment one; pa8/9, parapodium of segment eight or nine; pl, palp. Scale bars, 200 µm (**j**), 1 mm (**h**, **i**) and 2 mm (**a**, **b**, **d**, **e**; the scale bar in **e** also applies to **g**).

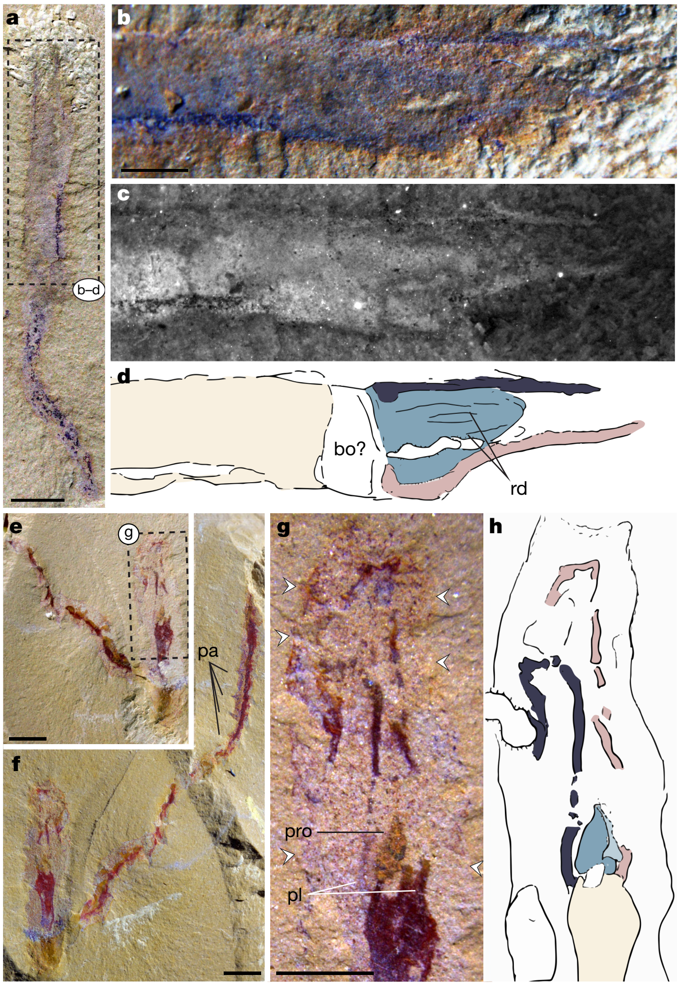


Fig. 2 | Anterior region of *D. tucolus*. **a**, YKLP 11390b, anterior fragment with head and palps. **b**, Anterior region of YKLP 11390b, low-angle illumination from the northwest. **c**, Region as in **b**, obtained using fluorescent light. **d**, Interpretative drawing outlining the anatomical features (colour scheme as in Fig. 1f). **e**, YKLP 11393b specimen, preserving the anterior region, dwelling tubes and a partial abdomen. **f**, YKLP 11393a, part. **g**, YKLP 11393b, details of the anterior region. The white arrowheads indicate the tube margins. **h**, Interpretative drawing of **g**. bo?, possible burrowing organ; pro, prostomium; rd, prostomial ridge. Scale bars, 500 µm (**b**), 2 mm (**a**, **g**) and 3 mm (**e**, **f**).

using fluorescence microscopy (Fig. 3d). The chaetae are directed slightly anteriorly (Fig. 1i); this feature is used to orient fragmentary specimens. In two specimens, lateral lamellae occur adjacent to chaetal bundles (Fig. 3g). Lamellae are crescent-shaped and approximately half the body width in length, with a dorsolateral (Extended Data Fig. 1k) to dorsal (Extended Data Fig. 5f–h) placement. As the rami are often parallel, a biramous morphology is revealed by subparallel chaetal bundles (Fig. 3d) or rare oblique views (Fig. 3h and Extended Data Fig. 5). Fine details of the chaetal morphology are obscure but are consistent with capillaries in the abdomen (Fig. 3d).

The pygidium is never well-preserved, but one specimen has putative pygidial cirri (Extended Data Fig. 5j, k). An ovoid structure between chaetiger one and three is of uncertain identity but resembles a blood lacuna (Fig. 1d–f and Extended Data Figs. 1h–j, 4g).

Eight specimens are preserved within a structure that extends beyond the body margin and chaetae (Fig. 3a–f, i–k and Extended Data Figs. 2–4). This structure is parallel to the body axis, approximately four times the body width (excluding parapodia) (Fig. 3a) with a sharp boundary with the matrix, which is visible using light (Fig. 2f) and fluorescence (Fig. 3h) microscopy and elemental maps for iron (Extended Data Fig. 3f). The structure is consistent with a dwelling tube, or tube-lined burrow. The tube has a slightly darker appearance relative to the matrix and lacks identifiable agglutinated bioclasts or grains. Tubes have a slight relief (Extended Data Fig. 3g) and sometimes have thick walls at their margins (Extended Data Fig. 6c–e), owing to compaction. As with the body, the tube contains iron (Fig. 3f) localized to small grains in the matrix that appear bright in backscatter images obtained using scanning electron microscopy (SEM) (Extended Data Fig. 3f). This is consistent with an organic composition, with organic material acting as a substrate for pyrite formation[8]. Specimens preserved in tubes vary from well-preserved with delineated parapodia and/or chaetae (Fig. 3c–e and Extended Data Fig. 3) to highly effaced, indicating in situ decay[9] (Extended Data Figs. 4, 6). Preservation quality varies along individual specimens (Extended Data Fig. 3a–d).

Discussion

Owing to low preservation potential[9], annelid body fossils are rare and distributed discontinuously and unevenly through geological time[10]. Diverse fossil polychaetes are known from early and middle Cambrian deposits (for example, Sirius Passet[3,11] and Burgess Shale[12]), but rare from China[13,14]. Cambrian annelids typically are generalized polychaetes in morphology, with well-developed biramous parapodia (suggesting motility[15]), elongated chaetae and a pair of palps[2,3,6,11]. They cannot be assigned to extant annelid subclades[6,12] and lack proposed annelid synapomorphies (for example, the prostomium–peristomium head structure[3] and typically pygidial cirri[12]) and are interpreted as stem-group annelids[6]. The lack of Cambrian crown annelids has

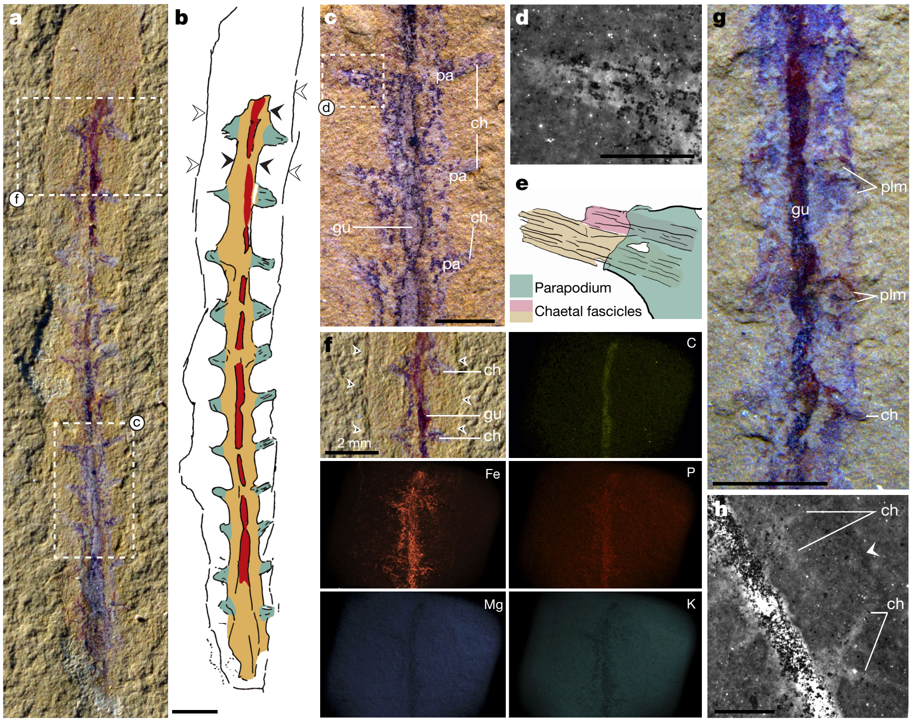


Fig. 3 | Morphological details of *D. tucolus*. **a**–**f**, YKLP 11383b. **a**, Abdominal region within a dwelling tube. **b**, Interpretative drawing of **a** (colours as in Fig. 1f). Anteriorly, black arrowheads indicate the body margin and white arrowheads indicate the tube edge. **c**, Magnification showing the segmented body and parapodia. **d**, Magnification of the chaetal fascicle. **e**, Interpretative drawing outlining the chaetal fascicles. **f**, Magnification and SEM–EDX maps. **g**, Magnification of YKLP 11389a showing the parapodia and lateral lamellae. **h**, Magnification of YKLP 11384a showing the biramous parapodia and chaetae. The white arrowhead indicates the tube edge. plm, parapodial lamella. Scale bars, 500 µm (**d**), 1 mm (**c**, **h**), 2 mm (**f**, **g**) and 4 mm (**b**).

prompted hypotheses of relatively late crown-group diversification, perhaps during the late Cambrian to Ordovician period[4], when jawed polychaetes become diverse and abundant[16].

Both molecular[1,5,17] and morphological[18] phylogenies have converged on scenarios in which annelids evolved from polychaete-like ancestors. However, reconciling fossil and phylogenomic evidence has been challenging[2,6,12,15]. Molecular phylogenies recover a grade of infaunal, sessile and tube-dwelling taxa as deep branches[1,5], including Magelonidae, Oweniidae, Chaetopteriformia and Sipuncula. These groups differ from Cambrian polychaetes in terms of gross morphology and inferred mode of life[6]. A tube-dwelling annelid ancestor has previously been proposed[19] (although see ref. [20]), which is contradicted by interpretations of the fossil record[15], and morphological hypotheses regarding the origin of annelid body plan features, for example, segmentation and parapodia[18,20]. Fossil specimens interpreted as sipunculans (which phylogenomic studies recover within Annelida[1,5]) are known from the early Cambrian Chengjiang Lagerstätte[21] but are rare and poorly known. If correctly interpreted, the oldest fossil crown annelids therefore belong to taxa that have lost most annelid synapomorphies[6], including segmentation[22]. Magelonidae and Oweniidae (Palaeoannelida[1]) are recovered as the sister group of all other annelids and so have featured prominently in recent discussions of the annelid ancestor[15,17,19]. These families are unusual among polychaetes, as they lack nuchal organs, possess monociliated epidermal cells and simple nervous systems[1,17].

Dannychaeta is dissimilar in gross morphology to that of previously known Cambrian polychaetes, but shares derived characters with extant Magelonidae. A spade-shaped prostomium with ventrolateral palps (Fig. 2) is characteristic of Magelonidae, which also have a differentiated thorax composed of eight or nine chaetigers[23,24]. Other well-known Cambrian annelids lack a clearly demarcated head and prostomial lobe[3] (but have lateral palps[15]), suggesting that the typical annelid head structure evolved after the origin of a segmented body, parapodia and palps[3]. The presence of a differentiated head in *Dannychaeta* is unique among Cambrian annelid fossils, indicating in itself that *Dannychaeta* has a phylogenetic position proximal to or within the annelid crown group[3]. The importance of the putative blood lacuna is uncertain, but small ring-shaped vessels occur in *Magelona*[23] and a larger lacuna occurs in *Poecilochaetus*[25]. However, in both species, these are placed more anteriorly, nearer the prostomium.

Although not widespread (around 7% of species[24]), several magelonid species live in tube-lined burrows[26] with an organic or parchment-like composition, some adhered sediment grains and/or bioclasts. These tubes are similar in inferred construction materials and dimensions relative to the body as in *Dannychaeta*. The presence of both obliquely and parallelly oriented specimens (Extended Data Figs. 2, 4) suggests at least some transport before burial.

We reconstruct *Dannychaeta* as a sessile, infaunal polychaete that fed in the water column using elongated palps (as in extant Magelonidae[24,26]) (Fig. 4a, b). Our phylogenetic analyses recover *Dannychaeta* in the magelonid stem group (Fig. 4c and Extended Data Figs. 7, 8). Parapodial and chaetal morphologies differ in some details in *Dannychaeta* and Magelonidae. In extant *Magelona*, the chaetae of the abdomen are hooded hooks, which are generally shorter than the thoracic capillary chaetae and occur in rows[24], whereas abdominal chaetae in *Dannychaeta* resemble capillaries held in bundles (Fig. 3c–e). Hooded hooks in some families share details of ultrastructure and formation[27], resulting in a proposed close relationship between Capitellidae, Spionidae and Magelonidae[27], which is not supported by phylogenomics[1]. The absence of hooded hooks in *Dannychaeta* may therefore provide consilient evidence of convergent chaetal evolution in these families. Parapodial lamellae also differ in some details in *Dannychaeta*, as they are dorsolaterally placed, but occur partially in inter-ramal space in some species in *Magelona*[24]. *Dannychaeta* is larger than extant magelonids that are typically less than 1 mm wide[24], although tubicolous species (for example, *Magelona alleni*) achieve the largest widths and share a more-robust anterior region (thorax) with *Dannychaeta*[24].

Other non-pleistoannelid polychaetes also share some characters with *Dannychaeta*. *Spiochaetopterus* has elongated palps, a differentiated anterior region and builds organic tubes[28]. However, chaetopterids have ridge-like parapodia (tori) with short, hooked chaetae (uncini[2]), which are distinct from laterally projecting, lobate parapodia in *Dannychaeta*. The head of *Spiochaetopterus* is neither a distinct anterior lobe nor spade-shaped. Nevertheless, chaetopterids branch proximally to Palaeoannelida[5]. A close relationship between *Dannychaeta* and chaetopterids would also suggest the presence of tubicolous, early branching crown-group annelids in the early Cambrian.

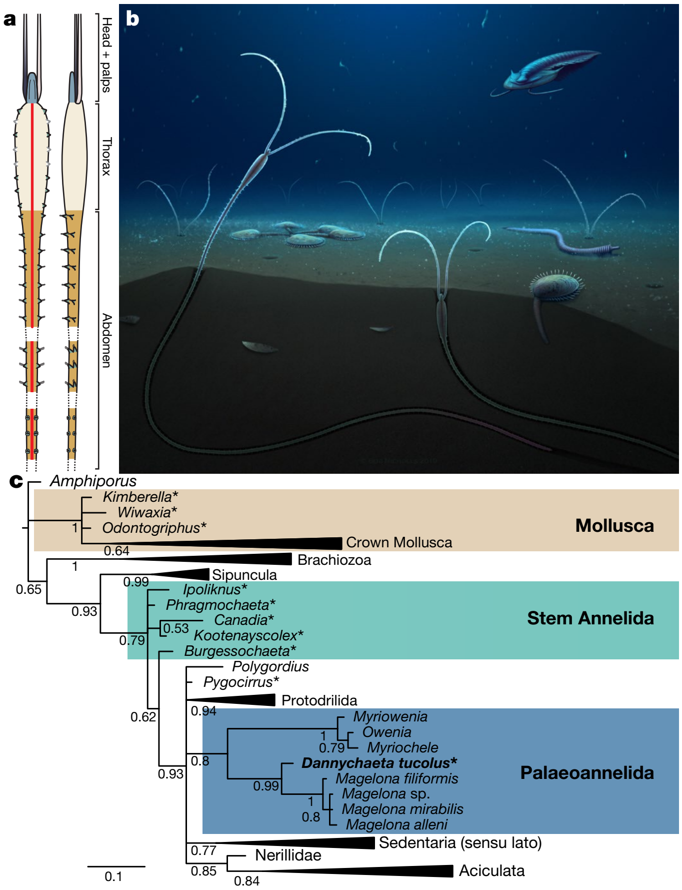


Fig. 4 | Reconstruction of *Dannychaeta*. **a**, Technical drawing showing proportions, body regions and gross anatomy. The colour scheme is as in Fig. 1f. The dorsal view (left) and right lateral view (right) are shown. Parapodia in the thorax that were not observed but inferred are shown in grey. **b**, Life reconstruction showing *D. tucolus* living in buried tubes. Artwork was created by R. Nicholls. **c**, Bayesian phylogenic analysis (365 characters, 143 taxa, mki + gamma model), incorporating *D. tucolus*. Numbers at the nodes are posterior probabilities; the scale bar indicates the number of substitutions per site (see Extended Data Figs. 7, 8 for full results and additional information). Fossil taxa are indicated by an asterisk.

Annelid tube fossils are well-documented[10]. Some late Ediacaran and early Cambrian fossil tubes are tentatively assigned to annelids or described as of 'annelid grade'[29,30], but lack diagnostic features that would enable confident phylogenetic placement. For example, the controversial late Ediacaran tubular fossil *Cloudina* has been tentatively reconstructed as an annelid based on the presence of a tubular gut[30]. Tube dwelling has evolved several times among extant annelids, and 'annelid-mimicking' taxa with lophophorate affinities are well-known from the early Palaeozoic era[29], indicating that caution should be exercised when assigning fossil tubes.

Regardless of the phylogenetic position and fossil record of sipunculans[4,21,22], *Dannychaeta* confirms that crown annelids are minimally early Cambrian in age, revealing early exploration of sessile ecological niches, more than a hundred million years before other unambiguous examples of tubicolous annelids[10,29]. *Dannychaeta* reveals that stem-group annelids coexisted with members of the crown group in the early Cambrian and exhibited a diversity of life modes, including epibenthic[15] and sessile forms.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2384-8.

1. Weigert, A. & Bleidorn, C. Current status of annelid phylogeny. *Org. Divers. Evol.* **16**, 345–362 (2016).
2. Eibye-Jacobsen, D. & Vinther, J. Reconstructing the ancestral annelid. *J. Zool. Syst. Evol. Res.* **50**, 85–87 (2012).
3. Parry, L., Vinther, J. & Edgecombe, G. D. Cambrian stem-group annelids and a metameric origin of the annelid head. *Biol. Lett.* **11**, 20150763 (2015).
4. Sperling, E. A. et al. MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Proc. R. Soc. B* **276**, 4315–4322 (2009).
5. Weigert, A. et al. Illuminating the base of the annelid tree using transcriptomics. *Mol. Biol. Evol.* **31**, 1391–1401 (2014).
6. Parry, L. A., Edgecombe, G. D., Eibye-Jacobsen, D. & Vinther, J. The impact of fossil data on annelid phylogeny inferred from discrete morphological characters. *Proc. R. Soc. B* **283**, 20161378 (2016).
7. Zeng, H., Zhao, F., Yin, Z., Li, G. & Zhu, M. A Chengjiang-type fossil assemblage from the Hongjingshao Formation (Cambrian Stage 3) at Chenggong, Kunming, Yunnan. *Chin. Sci. Bull.* **59**, 3169–3175 (2014).
8. Gabbott, S. E., Hou, X.-G., Norry, M. J. & Siveter, D. J. Preservation of Early Cambrian animals of the Chengjiang biota. *Geology* **32**, 901–904 (2004).
9. Briggs, D. E. G. & Kear, A. J. Decay and preservation of polychaetes: taphonomic thresholds in soft-bodied organisms. *Paleobiology* **19**, 107–135 (1993).
10. Parry, L. A., Eriksson, M. & Vinther, J. in *Handbook of Zoology: Annelida. Volume 1: Annelida Basal Groups and Pleistoannelida, Sedentaria I* (eds Purschke, G. et al.) 69–88 (De Gruyter, 2019).
11. Vinther, J., Eibye-Jacobsen, D. & Harper, D. A. An Early Cambrian stem polychaete with pygidial cirri. *Biol. Lett.* **7**, 929–932 (2011).
12. Eibye-Jacobsen, D. A reevaluation of *Wiwaxia* and the polychaetes of the Burgess Shale. *Lethaia* **37**, 317–335 (2004).
13. Liu, J. et al. Lower Cambrian polychaete from China sheds light on early annelid evolution. *Sci. Nat.* **102**, 34 (2015).
14. Han, J., Conway Morris, S., Hoyal Cuthill, J. F. & Shu, D. Sclerite-bearing annelids from the lower Cambrian of South China. *Sci. Rep.* **9**, 4955 (2019).
15. Parry, L. & Caron, J.-B. *Canadia spinosa* and the early evolution of the annelid nervous system. *Sci. Adv.* **5**, eaax5858 (2019).
16. Parry, L. A., Edgecombe, G. D., Sykes, D. & Vinther, J. Jaw elements in *Plumulites bengtsoni* confirm that machaeridians are extinct armoured scaleworms. *Proc. R. Soc. B* **286**, 20191247 (2019).
17. Helm, C. et al. Convergent evolution of the ladder-like ventral nerve cord in Annelida. *Front. Zool.* **15**, 36 (2018).
18. Westheide, W. The direction of evolution within the Polychaeta. *J. Nat. Hist.* **31**, 1–15 (1997).
19. Beckers, P. et al. The central nervous system of Oweniidae (Annelida) and its implications for the structure of the ancestral annelid brain. *Front. Zool.* **16**, 6 (2019).
20. Struck, T. Direction of evolution within Annelida and the definition of Pleistoannelida. *J. Zool. Syst. Evol. Res.* **49**, 340–345 (2011).
21. Huang, D. Y., Chen, J. Y., Vannier, J. & Saiz Salinas, J. I. Early Cambrian sipunculan worms from southwest China. *Proc. R. Soc. Lond. B* **271**, 1671–1676 (2004).
22. Carrillo-Baltodano, A. M., Boyle, M. J., Rice, M. E. & Meyer, N. P. Developmental architecture of the nervous system in *Themiste lageniformis* (Sipuncula): new evidence from confocal laser scanning microscopy and gene expression. *J. Morphol.* **280**, 1628–1650 (2019).
23. Jones, M. L. On the morphology, feeding, and behavior of *Magelona* sp. *Biol. Bull.* **134**, 272–297 (1968).
24. Mortimer, K. in *Handbook of Zoology* (eds Westheide, W. & Purschke, G.) (De Gruyter, 2019).
25. Allen, E. J. The anatomy of *Poecilochaetus*, Claparede. *Q. J. Microsc. Sci.* **48**, 79–151 (1904).
26. Mills, K. & Mortimer, K. Observations on the tubicolous annelid *Magelona alleni* (Magelonidae), with discussions on the relationship between morphology and behaviour of European magelonids. *J. Mar. Biol. Assoc. U. K.* **99**, 715–727 (2019).
27. Hausen, H. Chaetae and chaetogenesis in polychaetes (Annelida). *Hydrobiologia* **535/536**, 37–52 (2005).
28. Barnes, R. D. Tube-building and feeding in the chaetopterid polychaete, *Spiochaetopterus oculatus*. *Biol. Bull.* **127**, 397–412 (1964).
29. Ippolitov, A., Vinn, O., Kupriyanova, E. K. & Jäger, M. Written in stone: history of serpulid polychaetes through time. *Mem. Mus. Vic.* **71**, 123–159 (2014).
30. Schiffbauer, J. D. et al. Discovery of bilaterian-type through-guts in cloudinomorphs from the terminal Ediacaran Period. *Nat. Commun.* **11**, 205 (2020).

Methods

The specimens were studied, photographed at the Yunnan Key Laboratory for Palaeobiology, Yunnan University (YKLP) and at the University of Exeter, and are deposited at the YKLP. Photographs were taken using a Canon EOS 5DSR coupled with a MACRO 100-mm lens, and a Leica DFC7000T linked to a Leica M205 FA fluorescence microscope. The excitation wavelength of GFPL is 480 nm, and the excitation wavelength of RFPL is 546 nm. Images were obtained with a gain value of 3.3, saturation value of 52.00 and gamma value of 0.92. The external light source of the fluorescence microscope was a LEICA KL 300 LED, used for taking white-light images. SEM images were collected using a FEI Quanta 650 FEG using an accelerating voltage of 25 kV and a working distance of 12.4 mm. EDX analyses were carried out with an EDAX Pegasus using accelerating voltages of 25–30 kV with a working distance of 12.4–13 mm.

Phylogenetic analyses were based on a previously published character matrix for annelids and their close relatives[6], which has been updated successively with the addition of new taxa and fossil data[14–16,31]. We performed Bayesian analyses using MrBayes v.3.2.6 using the mki model with the Lewis correction for the scoring of only informative characters[32], with default priors for all parameters (that is, all trees were given equal prior probability). Bayesian analyses with and without topological constraints based on phylogenomic trees were used to investigate whether the conflicting topologies recovered from morphological and molecular data affected the phylogenetic position of *D. tucolus* (Extended Data Fig. 7). These constraints were constructed by incorporating results from the most recent[17] and taxon-rich[33,34] phylogenomic analyses of annelids and are outlined in detail in the Supplementary Information. We followed a previous study[15] by excluding *Arkonips* and *Guanshanchaeta* from the character matrix as they contain redundant character scores. For both analyses, 100 million generations were requested, with the analysis stopping once the average deviation of split frequencies dropped below 0.01. Convergence was then assessed using effective sample size (>200) and potential scale-reduction factor (about 1.0) values for all model parameters. Parsimony analyses without topological constraints were conducted in TNT 1.5[35] (courtesy of the Willi Hennig Society), using both equal weights and implied weights with $k = 10$. Bremer support and jack-knife and bootstrap frequencies from 1,000 replications were inferred for equal-weight trees and frequency differences were inferred from 1,000 replicates of symmetric resampling for the implied-weight trees.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data analysed in this paper are available as part of the Article, Extended Data Figs. 1–8 or Supplementary Information. The nomenclatural acts in this publication have been registered at ZooBank (LSID: urn:lsid:zoobank.org:pub:5BC89E47-2955-4539-94FD-D400E8C947FB).

Code availability

The phylogenetic dataset, and the commands and topological constraints necessary to run the MrBayes analyses, are included as NEXUS formatted files in the Supplementary Information.

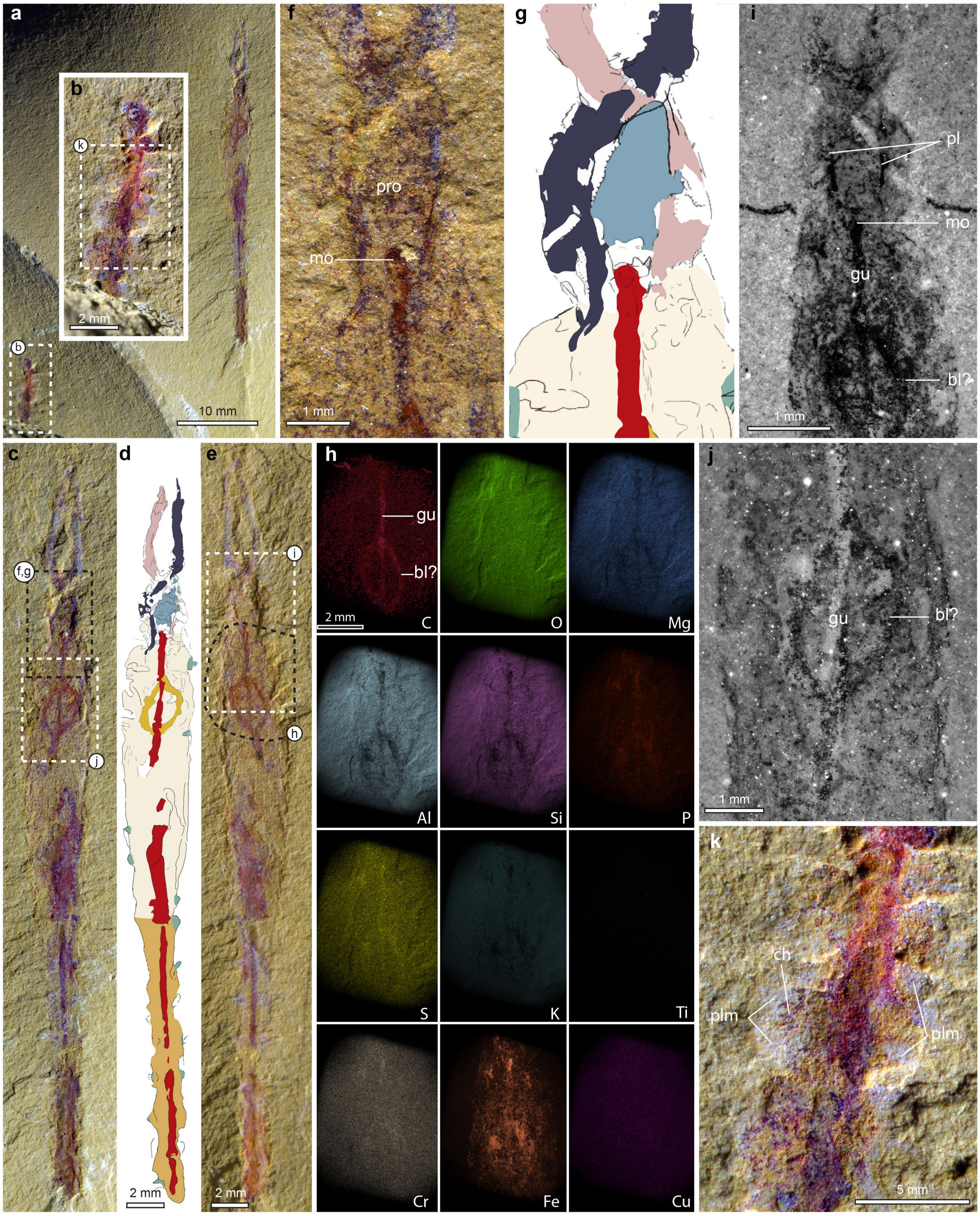
31. Nanglu, K. & Caron, J.-B. A new Burgess Shale polychaete and the origin of the annelid head revisited. *Curr. Biol*. **28**, 319–326 (2018).
32. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol*. **50**, 913–925 (2001).
33. Struck, T. H. et al. The evolution of annelids reveals two adaptive routes to the interstitial realm. *Curr. Biol*. **25**, 1993–1999 (2015).
34. Andrade, S. C. et al. Articulating "archiannelids": phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Mol. Biol. Evol*. **32**, 2860–2875 (2015).
35. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).

Acknowledgements This work was supported by NSFC grant 41861134032 and Yunnan Provincial grants 2015HA021, 2019DG050, 2015HC029 and 2018IA073. L.A.P. was supported by a YIBS Donolley Postdoctoral Fellowship. D.Z. is supported by the Key Research Program of the Institute of Geology & Geophysics, Chinese Academy of Sciences (IGGCAS-201905). X.M. is supported by a Natural Environment Research Council Independent Research Fellowship (NE/L011751/1). We thank K. Chen and colleagues from the Yuhua subdistrict office in Chenggong for support in field collection, P. Cong, D. Wu, T. Zhao, Y. Zhao and M. Yin for help with fieldwork, X. Yang and H. Mai for SEM–EDX technical support, R. Zhou for micro-computed tomography scanning, D. Eibye-Jacobsen for providing detailed comments on an earlier version of this manuscript and for his continued support and valuable mentorship, and R. Nicholls for the artistic reconstruction of *D. tucolus* shown in Fig. 4b.

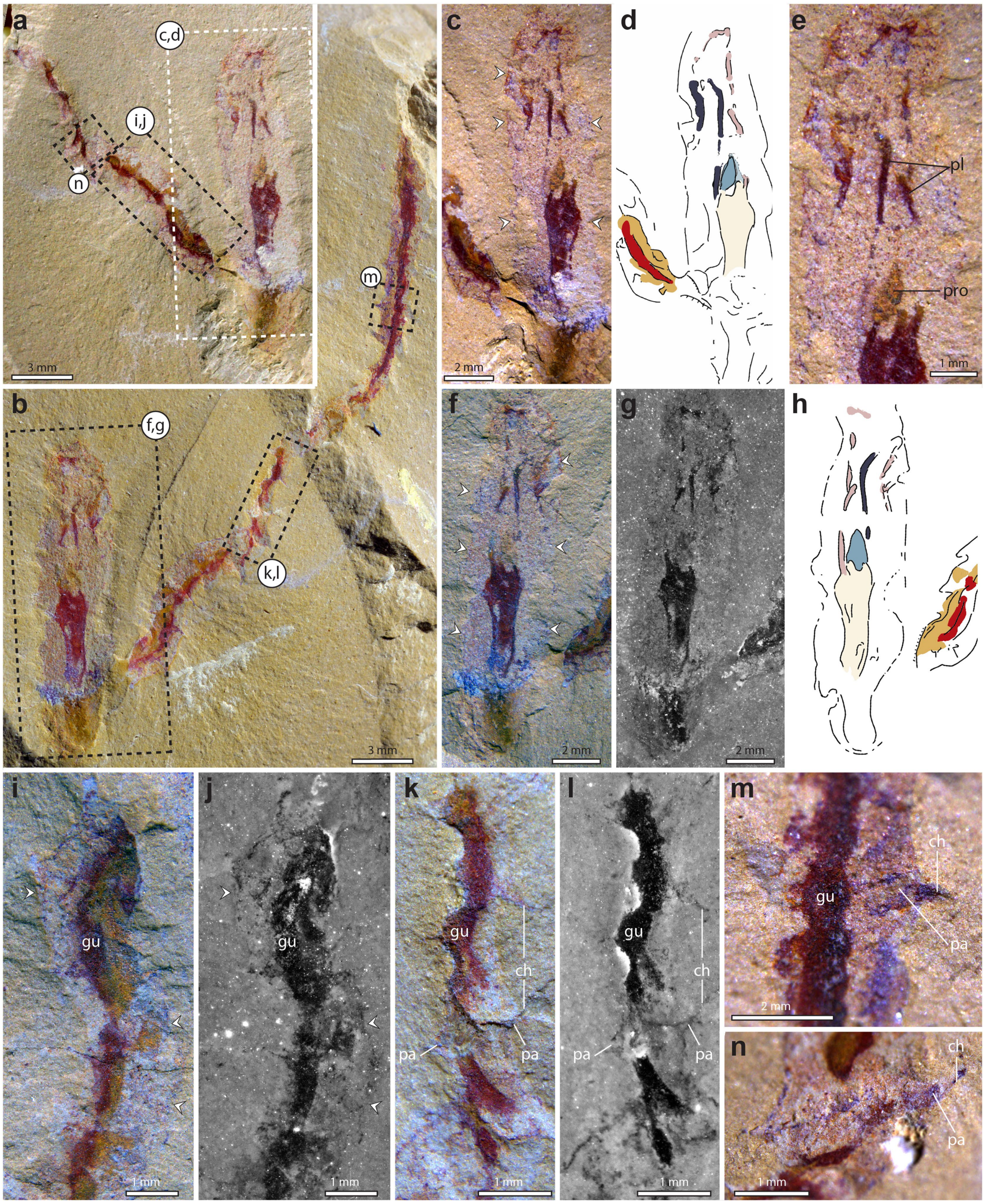
Author contributions L.A.P., H.C., J.V. and X.M. designed the study and interpreted the fossil specimens and their anatomy. H.C., D.Z. and X.H. collected the specimens. H.C. prepared and photographed all specimens and performed the EDAX elemental analysis. L.A.P. made the figures, performed the phylogenetic analyses and composed the first draft of the manuscript with substantial input from all co-authors.

Competing interests The authors declare no competing interests.

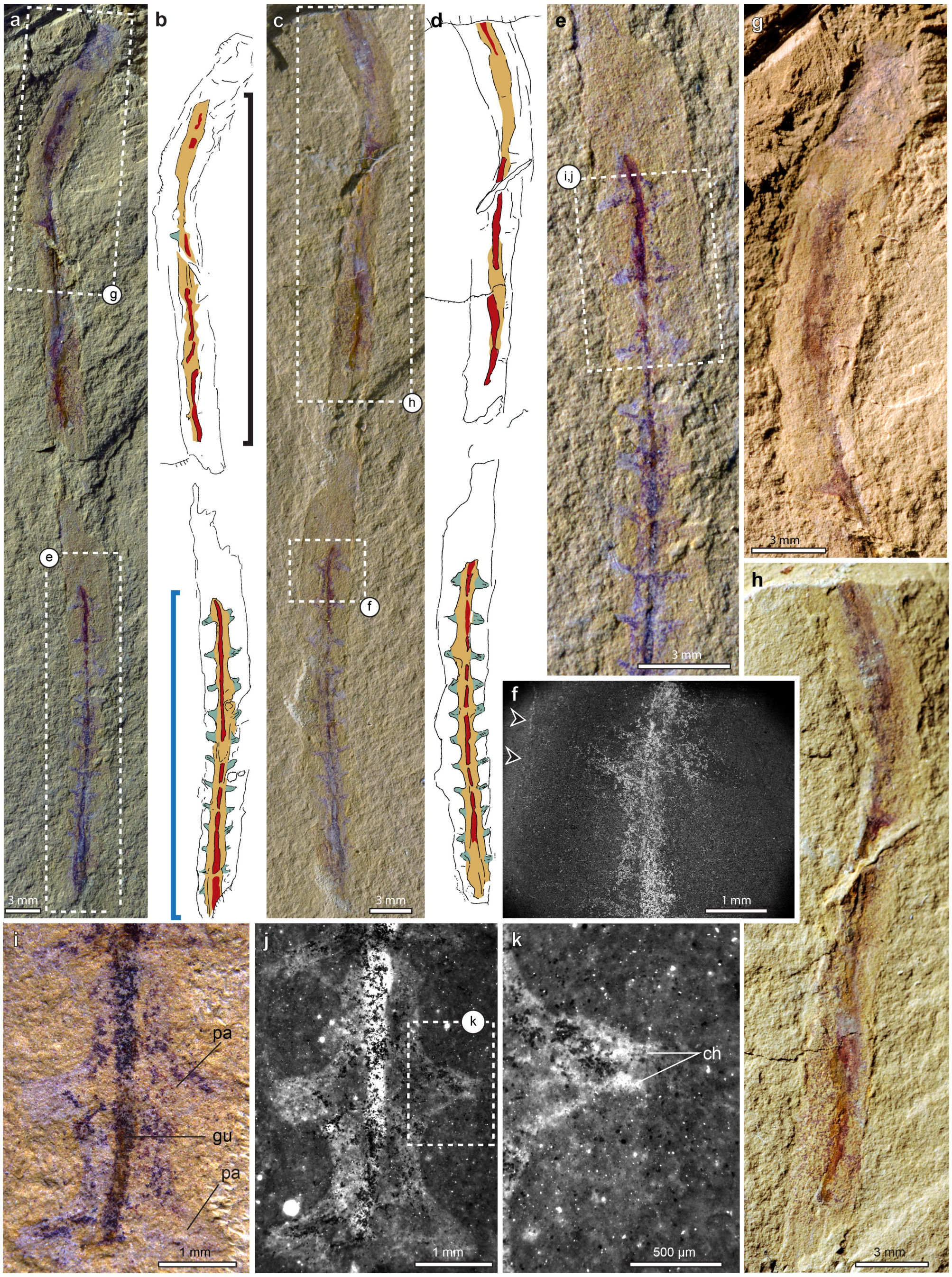
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2384-8.
Correspondence and requests for materials should be addressed to X.H. or X.M.
Peer review information *Nature* thanks Conrad Helm, Kate Mortimer, Mark D. Sutton and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



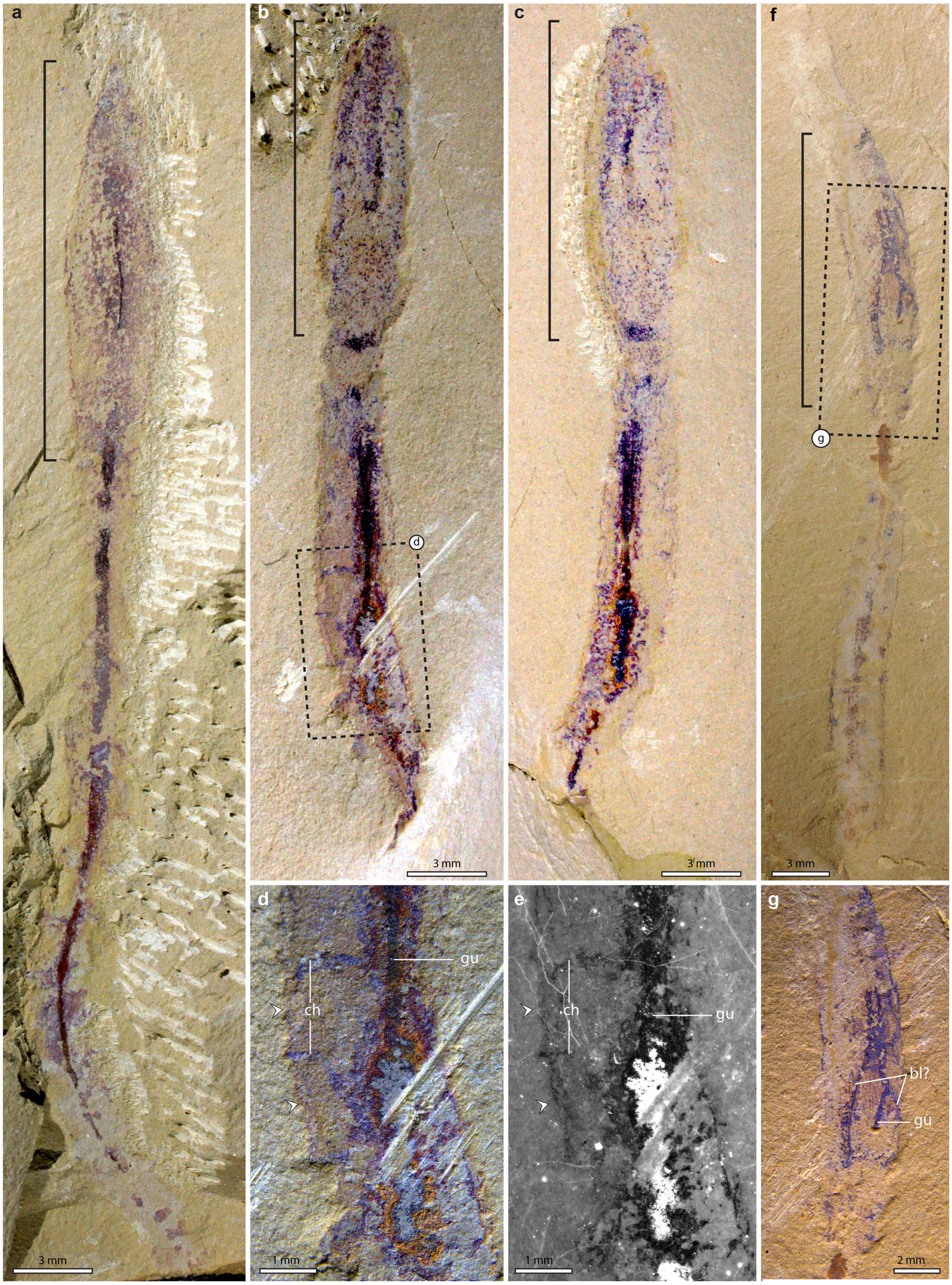
Extended Data Fig. 1 | Additional details of the holotype specimen. a, Entire specimen YKLP 11382a, including a possible posterior region. **b**, Possible posterior region from the region shown in **a**. **c**, YKLP 11382a, boxes show regions used for details of the head and interpretative drawing and fluoresence microscopy. **d**, Interpretative drawing of the specimen shown in **c**. **e**, YKLP 11382b. Image shows the regions used in SEM–EDX mapping and fluorescence microscopy. **f**, Magnification of the anterior region, showing the prostomial lobe and palp attachment. **g**, Interpretative drawing of the region shown in **f**, showing major anatomical features. The colour scheme is as in Fig. 1f. The palp attachment was inferred with additional information from the counterpart. **h**, SEM–EDX elemental maps. Element names for each map are shown in the bottom right corner. **i**, Fluorescence image showing the anterior region of the body in YKLP 11382b, including the gut, palps and putative blood lacuna. **j**, Fluorescence image of the putative blood lacuna in YKLP 11382a. **k**, Magnification of the posterior fragment associated with the holotype with parapodial lamellae. bl, blood lacuna.



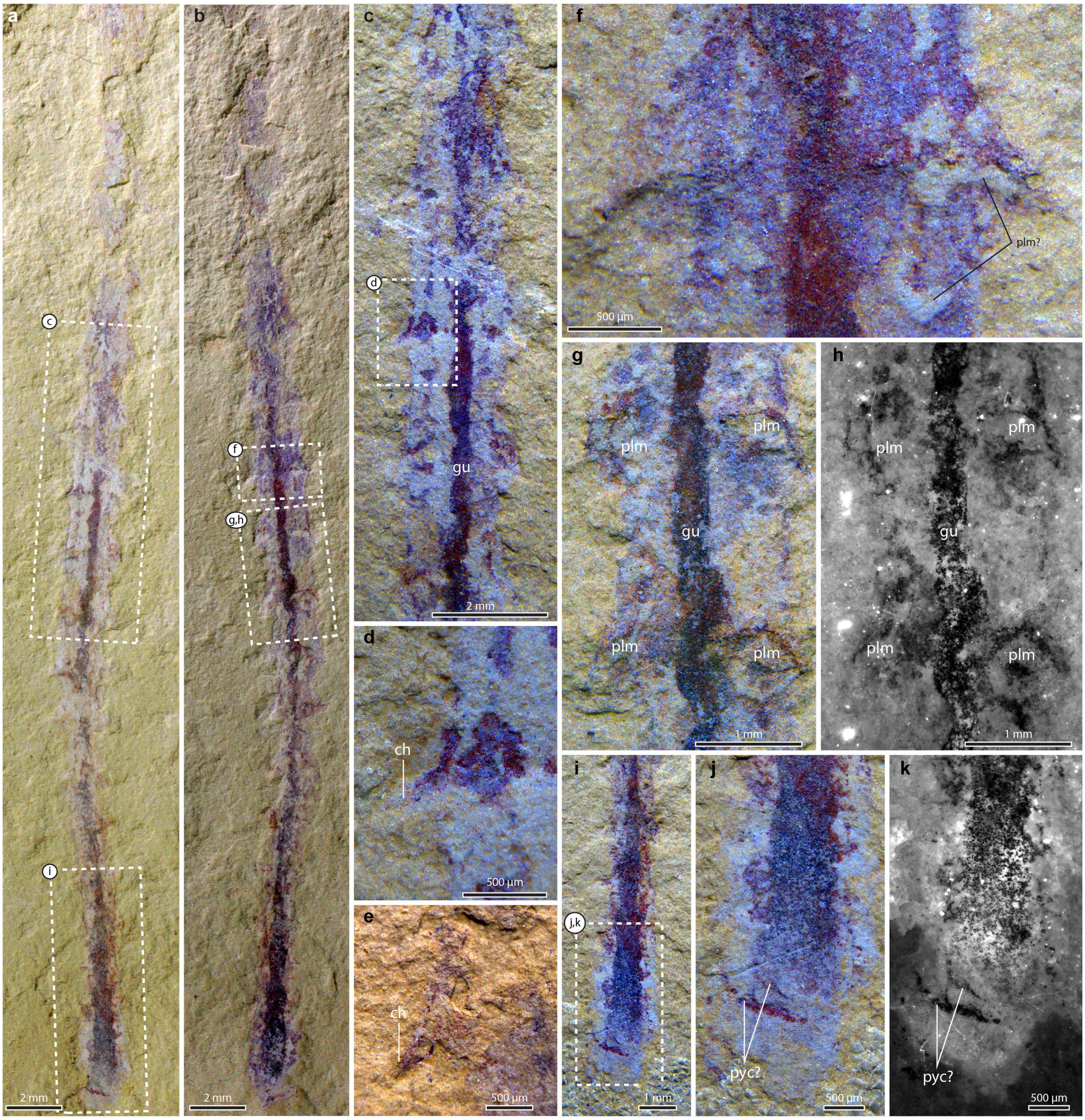
Extended Data Fig. 2 | Specimen of *D. tucolus* YKLP 11393 preserved in a dwelling tube. a, Specimen YKLP 11393b, for which the anterior region, dwelling tube and a partial abdomen are preserved. **b**, Specimen YKLP 11393a, for which the anterior region, dwelling tubes and a partial abdomen are preserved. **c**, Specimen YKLP 11393b. Magnification of the anterior region (boxed area shown in **a**). The white arrowheads indicate the tube margins. **d**, Interpretative drawing of the region shown in **c**. **e**, Magnification of the prostomium and palps shown in **c**. **f**, Specimen YKLP 11393a. Magnification of the anterior region (boxed area shown in **b**). The white arrowheads indicate the tube margins. **g**, Magnification of the same region as in **f**, imaged using fluorescence microscopy. **h**, Interpretative drawing of the region shown in **f** and **g**. **i**, Poorly preserved abdominal region, from the region shown in **a**, imaged using direct light. **j**, Same region as in **i**, imaged using fluorescence microscopy. **k**, Poorly preserved abdominal region, from the region shown in **b**, imaged using direct light. **l**, Same region as in **k**, imaged using fluorescence microscopy. **m**, Thoracic chaetiger showing parapodia and chaetae from the region shown in **b**. **n**, Thoracic chaetiger showing parapodia and chaetae from the region shown in **a**.



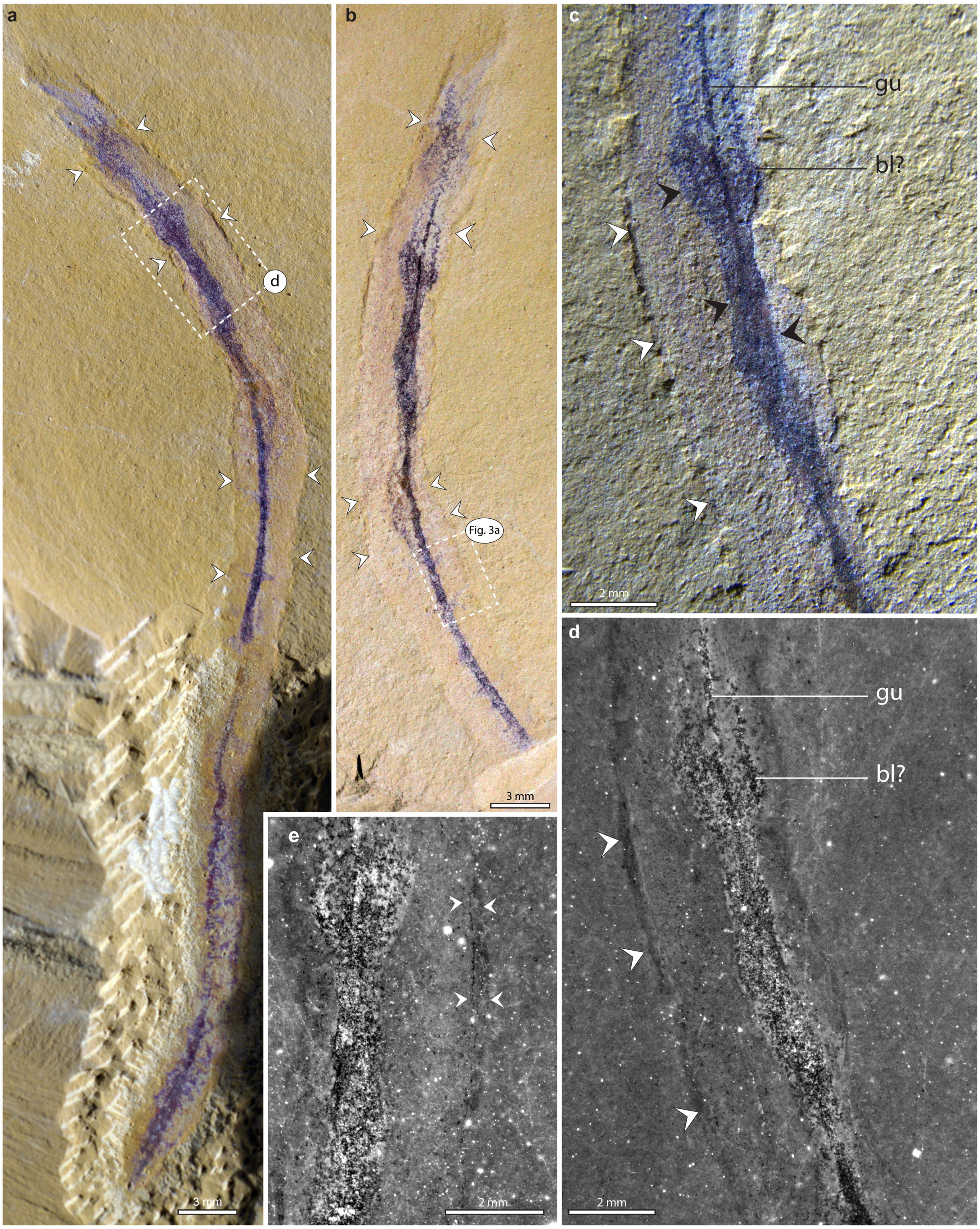
Extended Data Fig. 3 | Additional details of specimen YKLP 11383, preserved inside a dwelling tube parallel to the bedding. **a**, YKLP 11383a midbody fragment preserved inside a dwelling tube, part. **b**, Interpretative drawing of the specimen shown in **a**, regions demarcated by black and blue brackets represent decayed and well-preserved regions of the body fossil, respectively. **c**, YKLP 11383b, midbody fragment preserved inside a dwelling tube, counterpart. **d**, Interpretative drawing of the specimen shown in **c**. **e**, Magnification of a well-preserved region of YKLP 11383a as shown in **a**, showing 11 chaetigers preserved inside the dwelling tube. **f**, SEM backscatter image of a similar region to EDX maps shown in Fig. 3f, showing bright grains that are associated with the tube and body fossil. The arrowheads indicate the left tube margin and pyritized tube wall. **g**, Section of the fossil shown in **a** photographed under low-angle light to indicate the relief of the dwelling tube. **h**, Section of the fossil shown in **c** photographed under low-angle light to indicate the relief of the dwelling tube. **i**, Magnification of three chaetiger regions of the region shown in **e**. **j**, Same region as in **i**, photographed using fluorescence microscopy. **k**, Magnification of the individual parapodium shown in **j**, photographed using fluorescence microscopy.



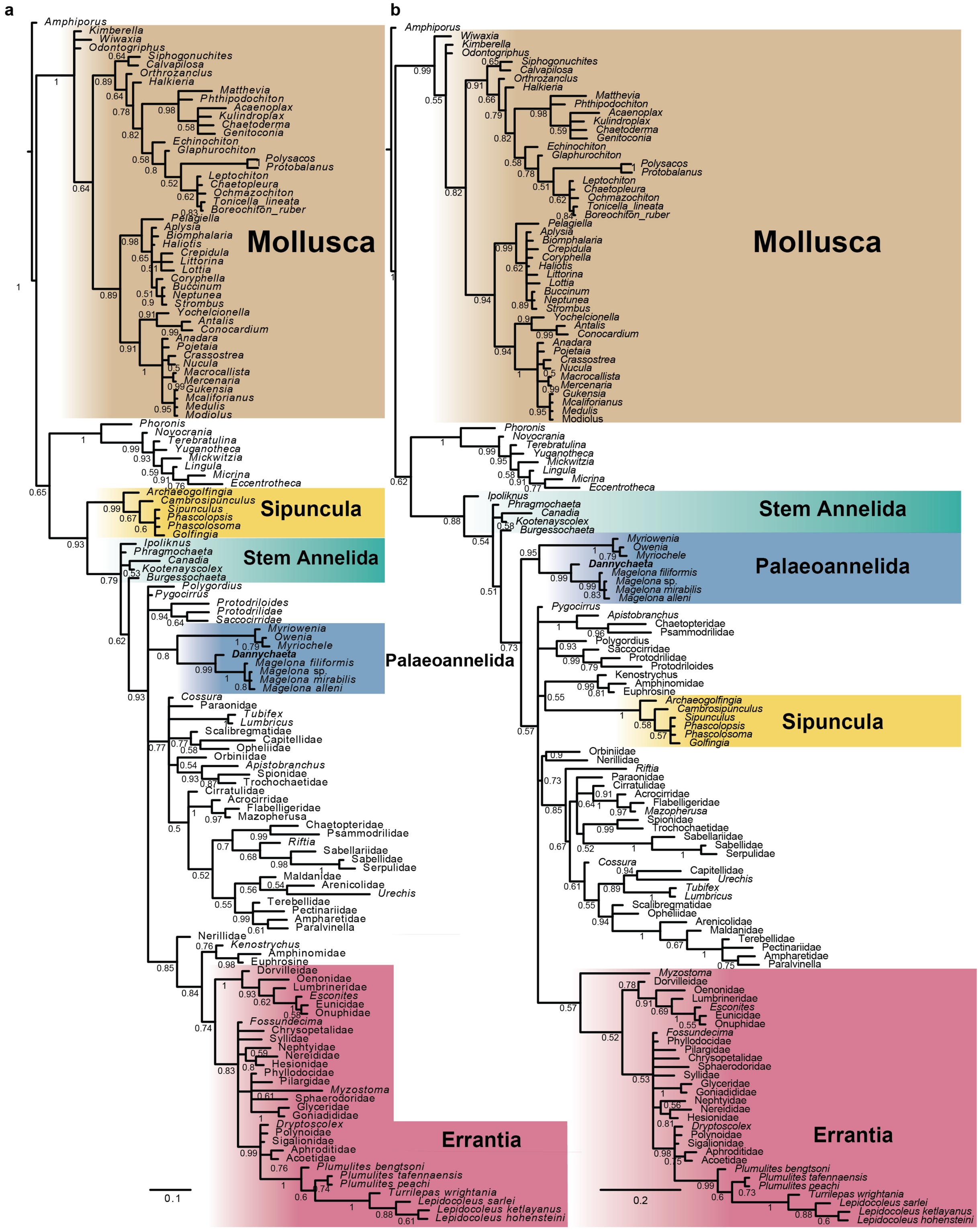
Extended Data Fig. 4 | Specimens YKLP 11385a, YKLP 11387 and YKLP 11401, showing effaced specimens preserved in dwelling tubes. **a**, YKLP 11385a, anterior fragment comprising the thorax and abdomen preserved inside dwelling tube. **b**, YKLP 11387a, anterior fragment preserving the thorax and abdomen. **c**, YKLP 11387b, anterior fragment preserving the thorax and abdomen. **d**, Magnification of the abdominal chaetigers in YKLP 11387a, from the region shown in **b**. The white filled arrows indicate the tube margins. **e**, Same region as in **d** imaged using fluorescence microscopy. **f**, YKLP 11401, effaced specimen preserved in a dwelling tube, including the putative blood lacuna. **g**, Magnification of the region shown in **f**, showing the gut and possible blood lacuna. Brackets in **a**–**c** and **f** indicate the position of the thoracic region.



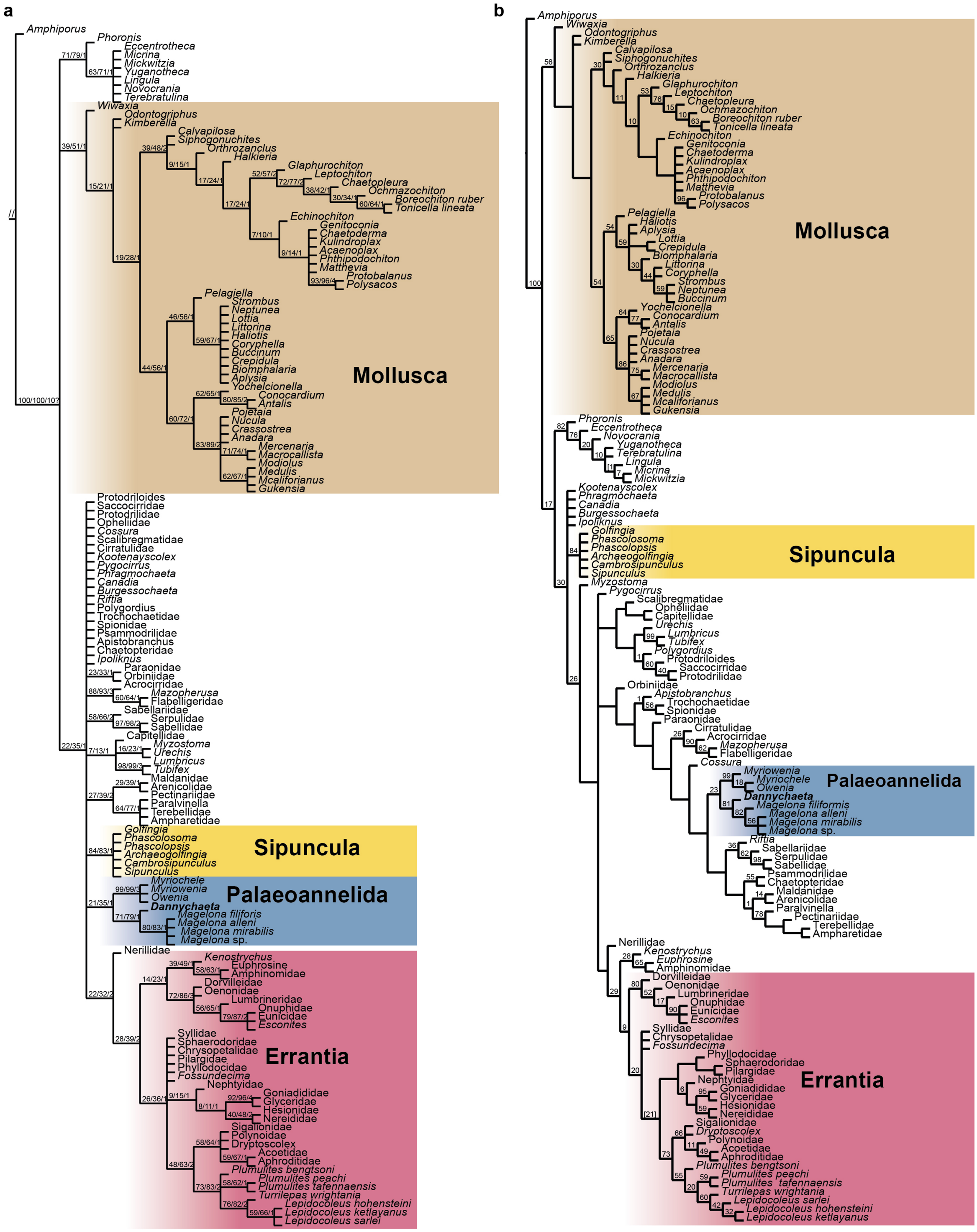
Extended Data Fig. 5 | Additional details of specimen YKLP 11389, showing details of parapodia, parapodial lamellae and the posterior region. **a**, YKLP 11389b, counterpart, posterior fragment preserving parapodia and chaetae. **b**, YKLP 11389a, part showing preservation of lateral parapodial lamellae. **c**, Magnification of five chaetigers from the region shown in **a**. **d**, Magnification of chaetiger from the region shown in **c**. **e**, Magnification of chaetiger in **d**. **f**, Magnification of chaetiger, preserving the parapodial lamellae from the region shown in **b**. **g**, Chaetigers preserving the parapodial lamellae from the region shown in **b**. **h**, Same region as in **g**, showing the parapodial lamellae, imaged using fluorescence microscopy. **i**, Posterior region as shown in **a**, with putative pygidial cirri. **j**, Magnification of putative pygidial cirri in **i**. **k**, same region as in **j**, imaged using fluorescence microscopy. plm, parapodial lamella; pyc, pygidial cirri.



Extended Data Fig. 6 | Specimen YKLP 11384, a decayed specimen preserved in a dwelling tube. **a**, YKLP 11384a (the part), a whole specimen in a dwelling tube. The white arrowheads indicate the tube margin. **b**, YKLP 11384b (the counterpart), a whole specimen in a dwelling tube. The white arrowheads indicate the tube margin. **c**, Details of YKLP 11384b, showing putative blood lacuna and gut. The white and black arrowheads indicate the tube and body margins, respectively. **d**, Same view as in **c**, but imaged using fluorescence microscopy. Note the thick appearance of the tube margin. **e**, Detail of YKLP 11384a, showing the preservation of the tube wall, indicated by white arrowheads.



Extended Data Fig. 7 | Full results of Bayesian phylogenetic analyses. a, Full results of unconstrained analysis under the mki + gamma model. The topology shows a majority rule consensus tree, the scale bar is in units of expected number of substitutions per site, the numbers at the nodes are posterior probabilities. **b**, Full results of an analysis with constraints from phylogenomics under the mki + gamma model. The topology shows a majority rule consensus tree, the scale bar is in units of expected number of substitutions per site and the numbers at the nodes are posterior probabilities.



Extended Data Fig. 8 | Results of parsimony analyses under equal and implied weighting. a, Strict consensus tree (length 1,071) of parsimony analysis without topological constraints, numbers at the nodes are support from bootstrapping, jack-knifing and Bremer decay. **b**, Strict consensus of trees inferred using implied weighting with $k = 10$ (tree score 46.14190), numbers at the nodes are relative frequencies from symmetric resampling.

nature research

Corresponding author(s): Xiaoya Ma and Xianguang Hou

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	☐	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection: the morphological character matrix was compiled using Mesquite 3.6 (https://www.mesquiteproject.org/)

Data analysis: Phylogenetic analyses using MrBayes 3.2.6 (open source) and TNT 1.5 (freely available from the Willi Hennig Society).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The morphological character matrix is supplied in the supplementary data and all other relevant data is included in the manuscript and extended data figures.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The material studied are fossil specimens collected from the early Cambrian strata in Southwest China, which have been imaged using various microscopy techniques, described and included in morphological phylogenetic analyses.
Research sample	Fossil specimens were collected from the early Cambrian Canglangpu Formation (circa 514Ma), Yunnan Province, China . The morphological matrix was based on a previously published analysis available from Parry et al (2019, PRSB).
Sampling strategy	No statistical methods were used to determine sample size
Data collection	Light photomicrographs were collected using Leica microscopes. SEM EDX images were collected using FEI Quanta 650 FEG.
Timing and spatial scale	N/A
Data exclusions	Two taxa (Arkonips and Guanshanchaeta) were excluded from phylogenetic analyses as they contain redundant information, following Parry and Caron (2019); Science Advances.
Reproducibility	N/A
Randomization	Specimens were determined to belong to the same, new species as they share characters not seen in other fossil taxa from the deposit or Cambrian fossil record in general.
Blinding	Blinding of data is not relevant to morphological study of fossil specimens or phylogenetic analysis

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The fossil specimens were collected from several outcrops in small hills around the Guanshan reservoir. The field collecting was carried out regularly between October 2015 to July 2016 by the several students and staff from the Yunnan Key Laboratory for Palaeiobiology (YKLP).
Location	Southwest of the Guanshan reservoir, Chenggong, Kunming (24° 49’ 37” 70 N, 102° 50’ 00” E), China.
Access and import/export	The fossil collecting activities have followed strictly the laws in China and in Yunnan Province. All collecting permissions were obtained before the collecting started. Any collecting in private land also obtained permissions from the land owner.
Disturbance	The fossil material have been collected from several outcrops with good sediment exposure, so little vegetation was damaged. The outcrops are small, and the collecting was carried out with geological hammers, so no machinery was applied.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☐ Unique biological materials
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☒ Palaeontology
☐	☐ Animals and other organisms
☐	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	N/A

Antibodies

Antibodies used	N/A
Validation	N/A

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	N/A
Authentication	N/A
Mycoplasma contamination	N/A
Commonly misidentified lines (See ICLAC register)	N/A

Palaeontology

Specimen provenance	All specimens, YKLP 11383 to YKLP 11402, were collected by authors and stored at the Yunnan Key Laboratory for Palaeobiology, Yunnan University.
Specimen deposition	Fossil material is accessioned to the collection of the Yunnan Key Laboratory for Paleobiology (YKLP), Yunnan University
Dating methods	N/A

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	N/A
Wild animals	N/A
Field-collected samples	N/A

Human research participants

Policy information about studies involving human research participants

Population characteristics	N/A
Recruitment	N/A

Structure and flexibility in cortical representations of odour space

https://doi.org/10.1038/s41586-020-2451-1

Received: 7 August 2018

Accepted: 2 April 2020

Stan L. Pashkovski[1], Giuliano Iurilli[1,2], David Brann[1], Daniel Chicharro[1,3], Kristen Drummey[1], Kevin Franks[4], Stefano Panzeri[3] & Sandeep Robert Datta[1]✉

The cortex organizes sensory information to enable discrimination and generalization[1–4]. As systematic representations of chemical odour space have not yet been described in the olfactory cortex, it remains unclear how odour relationships are encoded to place chemically distinct but similar odours, such as lemon and orange, into perceptual categories, such as citrus[5–7]. Here, by combining chemoinformatics and multiphoton imaging in the mouse, we show that both the piriform cortex and its sensory inputs from the olfactory bulb represent chemical odour relationships through correlated patterns of activity. However, cortical odour codes differ from those in the bulb: cortex more strongly clusters together representations for related odours, selectively rewrites pairwise odour relationships, and better matches odour perception. The bulb-to-cortex transformation depends on the associative network originating within the piriform cortex, and can be reshaped by passive odour experience. Thus, cortex actively builds a structured representation of chemical odour space that highlights odour relationships; this representation is similar across individuals but remains plastic, suggesting a means through which the olfactory system can assign related odour cues to common and yet personalized percepts.

In olfaction, perception depends on chemistry[8]. Chemically related odours evoke similar percepts within and across individuals, suggesting that the cortex harbours a conserved mapping from chemical to neural space that organizes information about odour relationships to ultimately support perception[6,7]. Odours are detected by broadly tuned receptors expressed by olfactory sensory neurons, the axons of which project to the olfactory bulb (OB)[9,10]. Within the mouse OB, these axons are organized into thousands of discrete and spatially organized information channels known as glomeruli, each of which represents the tuning properties of an individual odour receptor[11]. Odour information is reformatted by OB circuits before being transmitted to cortex; it is not clear whether or to what degree this peripheral transformation preserves information about odour chemical relationships[12–14].

The main recipient of OB afferents is the piriform cortex (PCx)[1]; axons from OB projection neurons are broadly dispersed across the entire surface of the PCx, and individual PCx neurons respond to multiple, chemically distinct odorants[15–17]. These observations suggest that neurons in the PCx randomly sample sensory inputs from the OB[18,19]. Consistent with this possibility, individual odours activate ensembles of spatially distributed PCx neurons that lack apparent topographical organization with respect to chemical space[18,20,21]. Feed-forward random network models (which posit stochastic connectivity between OB glomeruli and PCx neurons) predict that PCx odour representations should be pervasively decorrelated, but that PCx should maintain the pairwise odour relationships present in the OB; these models further suggest that cortical codes for odour relationships should be invariant across individuals, as peripheral representations of chemical relationships are largely determined by the tuning properties of odour receptors, which are encoded in the genome[19,22–24].

However, in addition to receiving inputs from the OB, PCx neurons are linked through a dense web of excitatory interconnections, which suggests that the olfactory cortex acts as an auto-associative network[1,25]. Such networks use Hebbian mechanisms to construct cell assemblies that encode information about stimulus relationships (such as feature similarity or temporal coincidence) through correlated activity. In the case of PCx, auto-associative mechanisms are predicted to both increase generalization across chemically similar odours, and to render cortical odour representations sensitive to passive odour experience, thereby reshaping pairwise odour relationships inherited from OB inputs. Although the PCx exhibits characteristics that are consistent with both random and auto-associative networks, it remains unclear whether the cortex systematically encodes information about odour chemical relationships; whether any such representation preserves or reshapes odour relational information conveyed by the OB; or whether cortical odour representations are primarily decorrelated (thereby favouring odour discrimination as predicted by random network models) or correlated (thereby favouring odour generalization as predicted by auto-associative models).

Cortex encodes odour chemical relationships

To address these questions, we used multiphoton microscopy in mice expressing the fluorescent Ca^{2+} indicator GCaMP6s within the PCx to assess neural activity both in the input-dominated PCx layer 2 (L2), and

---

[1]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [2]Center for Neuroscience and Cognitive Systems, Istituto Italiano di Tecnologia, Rovereto, Italy. [3]Neural Computation Laboratory, Istituto Italiano di Tecnologia, Rovereto, Italy. [4]Department of Neurobiology, Duke University, Durham, NC, USA. ✉e-mail: srdatta@hms.harvard.edu

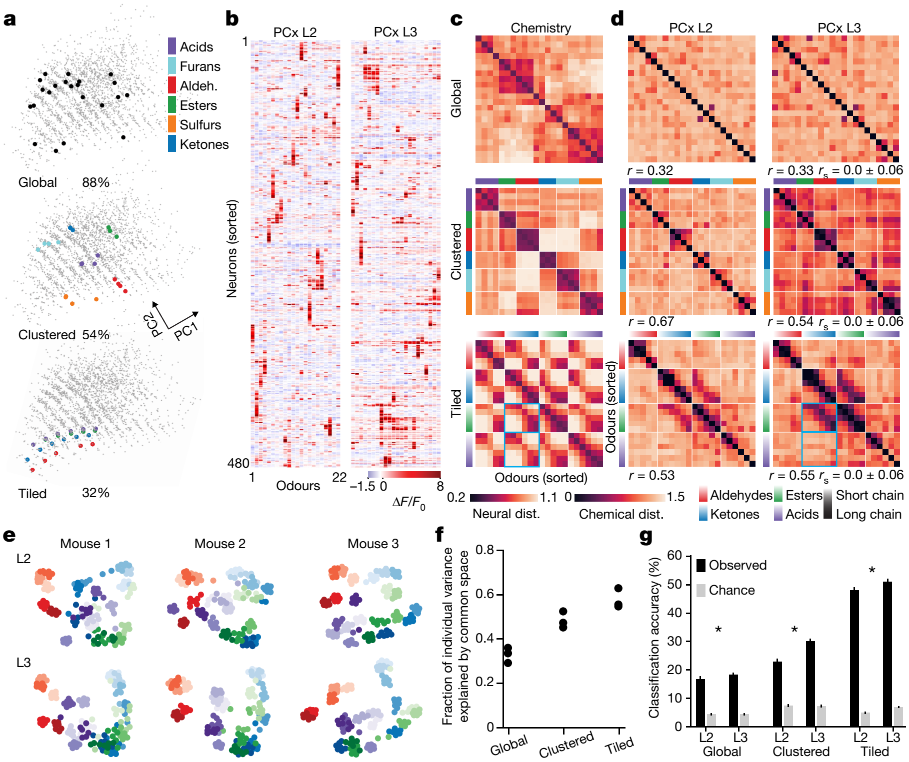


Fig. 1 | Systematically probing relationships between odour chemistry and cortical odour representations. **a**, Global, clustered and tiled odour sets (see Extended Data Fig. 1e for odour identities and structures), depicted in principal component space (see Methods). Colour indicates functional group associated with each odour. The amount of variance spanned by each odour set (of the full odour space, grey dots) is indicated. **b**, Example single neuron responses for the clustered odour set, representing the trial-averaged response of single neurons (rows) across 22 odours (columns). Rows are sorted using hierarchical clustering, with PCx L2 and L3 rasters sorted independently (Methods). **c**, Pairwise odour distances (Pearson's correlation) for all odour sets based on chemical descriptors (Methods). Rows and columns represent individual odours sorted using hierarchical clustering (ordering as in Extended Data Fig. 1e). Colour bars indicate functional groups associated with each odour. **d**, Pairwise odour distances based on pooled neural population responses in PCx L2 and L3 (Methods), sorted as in **c**. Pearson's correlation coefficient between the chemical and neural distance matrices reported below each matrix (global: $P < 10^{-7}$; clustered: $P < 10^{-16}$; tiled: $P < 10^{-18}$); r_s (shuffle) obtained by independently permuting odour labels for each neuron. Blue boxes highlight ketone-ester and ketone-acid relationships between chemistry and PCx L3. **e**, UMAP embeddings of cortical responses to the tiled odour set. Each dot represents a population response for one odour presentation (7 per odour), colour-coded as in **d**. **f**, Fraction of total variance in each mouse (L3 activity) attributable to shared across-mouse structure determined by distance covariance analysis (Methods). **g**, *k*-nearest-neighbour classification of odour identity in a held-out mouse using odour distances from other mice. Data are bootstrap mean ± s.e.m.; grey bars indicate shuffle control on odour labels (Methods). (Accuracy is greater in PCx. global: $P < 10^{-3}$; clustered: $P < 10^{-60}$; tiled: $P < 10^{-22}$, two-sided Wilcoxon rank sum test.) Data in **b**, **d**–**g** are based on all responsive neurons (Methods) pooled by layer across mice (*n* mice, neurons (L2/L3) for global: 3, (854/616), clustered: 3, (867/488), tiled: 3, (427/334)) (see Methods for subject-specific statistics).

in the more associational layer 3 (L3, in which odour responses have not yet been described)[26] (Extended Data Fig. 1). We took advantage of a library of odour descriptors that quantifies thousands of physiochemical features, such as molecular weight, polarizability and hydrophobicity[5,27], to rationally design three sets of 22 odours each: a 'global' odour set, which included structurally diverse odorants well separated in odour space; a 'clustered' odour set divided into six odour subsets, each of which shared functional groups and other structural features; and a 'tiled' odour set, in which the carbon chain length of a ketone, an ester, an aldehyde and an acid was incrementally varied (Fig. 1a, Extended Data Fig. 1, Methods). Although each odour set captured progressively less chemical variance, by construction individual odours in the clustered set (within each of the six subsets) were most closely related, whereas odours were separated at intermediate distance scales in the tiled set. We noted that under anaesthesia odour responses in L3 (and to a lesser extent L2) were attenuated or absent; recordings were therefore performed during wakefulness, a state in which L3 neurons were considerably more active (Extended Data Fig. 2, Methods).

All odours evoked selective excitation and suppression, with PCx L3 responses being denser, broader and more reliable than those in L2 (Extended Data Fig. 3). Odours evoked more correlated activity across the population of PCx neurons (that is, ensemble correlations) than was expected by chance, with greater correlations observed in L3 compared to L2 (Fig. 1b, Extended Data Fig. 3). These findings raised the possibility that correlated odour-evoked responses among PCx ensembles systematically reflect chemical relationships among odour stimuli. To explore this possibility, correlation distance matrices were generated for each odour set based on the physiochemical descriptors that characterize each odorant (Fig. 1c, Methods). Odours in the global set were the least chemically correlated with each other, whereas odours in the clustered odour set exhibited substantial block diagonal structure, consistent with subsets of odours sharing key chemical attributes. Because molecules in the tiled set are related along two chemical axes (for example, heptanone and octanone differ by one carbon atom, whereas heptanone and pentyl acetate differ by one oxygen atom), the matrix describing these odours exhibited periodic on- and off-diagonal structure.

Visual comparison and quantification demonstrated that odour chemistry and neural responses were only weakly related in the global odour set; by contrast, cortical odour responses maintained the block

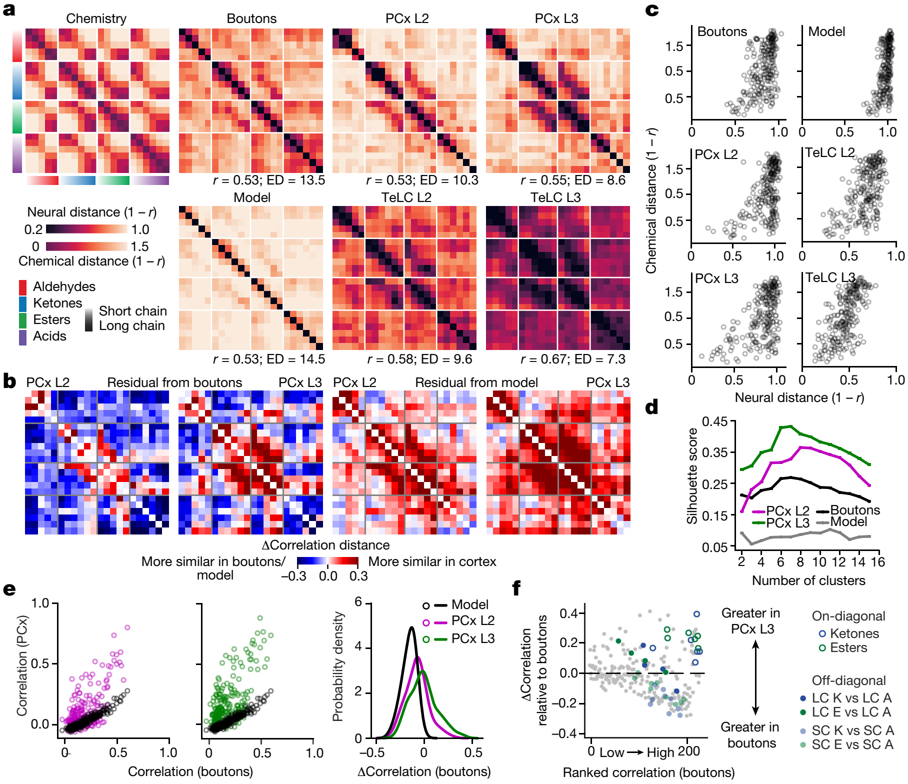


Fig. 2 | Correlation structure differs in olfactory bulb and cortex. a, Correlation distance matrices for the tiled odour set across all conditions. Top left, distances obtained using chemical descriptors. Right, distances based on odour responses. Odour sorting as in Fig. 1c. *r* values indicate Pearson's correlation with odour chemistry (Boutons: $P < 10^{-17}$; PCx L2: $P < 10^{-17}$; PCx L3: $P < 10^{-19}$; Model: $P < 10^{-17}$; TeLC L2: $P < 10^{-21}$; TeLC L3: $P < 10^{-32}$; shuffled Pearson's $r = 0.0 \pm 0.063$ (mean ± s.d.), 1,000 permutations on odour label). ED, effective dimensionality (Methods). **b**, Left, difference between PCx and bouton distances in **a**. Right, difference between PCx and random network model distances in **a** (Methods). **c**, Pairwise odour correlation distances based on neural responses plotted against corresponding chemical distances. **d**, Silhouette scores for clustered population responses (based upon Euclidean distances and grouped via *k*-means clustering) over a range of cluster sizes. Higher values indicate better clustering (Methods). **e**, Left, pairwise odour correlations in boutons and PCx predicted by the feed-forward random network model (Methods) compared to observed correlations in PCx L2 and L3. Right, probability density distribution of differences between cortical (PCx L2 and L3) and input (boutons) pairwise odour correlations, superimposed on the distribution expected with the model (model versus L3: $P < 10^{-33}$, versus L2: $P < 10^{-17}$, Kolmogorov–Smirnov test). **f**, Difference in pairwise odour correlations between PCx L3 and boutons (grey dots). Positive values indicate greater correlation in the cortex. Odour pairs are ranked along the *x* axis from least to highest correlation in the bouton data. Short-chain (SC) and long-chain (LC) comparisons between ketones (K), esters (E) and aldehydes (A) are colour-coded as shown.

diagonal physiochemical correlation structure apparent in the clustered odour set, demonstrating that at close chemical distances, PCx represents odour chemical relationships (Fig. 1d). Notably, neural responses to the tiled odour set (in which odour relationships are organized at intermediate chemical distances) reflected on-diagonal chemical relationships, but did not uniformly encode off-diagonal relationships. For example, the cortex appeared to emphasize chemical similarities between ketones and esters, while de-emphasizing chemical similarities between ketones and acids (Fig. 1d, highlighted blue boxes). Structured chemical–neural relationships were apparent on a trial-by-trial basis, and persisted for several seconds after odour offset; as has been observed previously under anaesthesia, no spatial ordering of neurons was observed with respect to odour chemistry during wakefulness, consistent with response correlations alone conveying information about odour relationships[18,20] (Extended Data Fig. 4).

Both uniform manifold approximation and projection (UMAP) embeddings and manifold alignment revealed that cortical odour relationships were similar across mice (Fig. 1e, f); indeed, information about pairwise cortical odour distances derived from one mouse could be used to predict the identity of a held-out odorant based upon odour distances measured in a different mouse, with better performance observed in L3 than L2 (Fig. 1g, Methods). Lasso optimization was used to identify chemical features relevant to driving neural responses in each of the odour sets; identified descriptors captured physiochemical features such as molecular weight, electronegativity, polarizability and hydrophobicity, which suggests that ensemble-level odour representations are driven by diverse aspects of odour chemistry (Supplementary Table 1, Methods). Identified features that predicted neural activity for each odour set also improved the correspondence between all the other odour sets and their associated neural activity, demonstrating that information about odour chemistry gleaned from one experiment can be used to predict cortical responses in a different experiment carried out using a separate set of odorants (Extended Data Fig. 5a).

Cortical odour representations reshape bulb inputs

The selective differences between odour chemical relationships and cortical activity apparent in the tiled odour experiment could reflect correlation structure present in OB inputs to PCx (consistent with feed-forward random network models), or instead could be generated by cortex (consistent with auto-associative models). However, until now it has not been possible to quantify odour-evoked responses across the complete array of OB glomeruli, which has prevented the characterization of correlation structure in bulb inputs to PCx. To address this challenge, we introduced synaptically targeted GCaMP6s into projection neurons spanning the OB, and imaged odour evoked activity in boutons in PCx layer 1a (L1a), where they synapse with L2 and L3 neurons; because the axons and boutons of all OB glomeruli are spatially distributed across the PCx[15,16], each cortical field of view effectively samples glomeruli from the entire bulb (Methods, Extended Data Fig. 6).

Odours from the tiled odour set evoked both excitation and suppression in OB boutons, the responses of which were similar across mice (Extended Data Fig. 7). Correlation distance matrices revealed that bouton responses reflected information about odour chemical relationships (Fig. 2a); in addition, identification of physiochemical features that optimized the observed chemical-bouton relationships improved predictions of bouton responses to held-out odours as well as predictions of cortical responses to the tiled odour set (Extended Data Fig. 5b, Supplementary Table 1). Thus, similar to the cortex, OB projection neuron boutons encode information about odour relationships and chemistry.

However, odour responses in boutons and cortex exhibited distinct patterns of correlation with respect to chemistry, with the greatest chemical–neural differences observed in PCx L3 (Fig. 2a, b). Although the average level of correlated activity was similar in boutons and cortex, the distribution of odour-evoked correlations differed, with bouton representations exhibiting higher effective dimensionality (see Methods); by contrast, odour responses of PCx L3 neurons were more clustered, more selectively structured, and exhibited both lower effective dimensionality and a wider dynamic range for representing close chemical relationships (Fig. 2a, c–e, Extended Data Fig. 7). The presence of these structured correlations in part reflected increased grouping of closely related odorants, as representations for odours nearest each other in chemical space (that is, on-diagonal correlation matrix relationships) were more clustered in the cortex than in boutons (Fig. 2a, f). One exception to this trend was acids, which as a class were correlated in the OB but relatively decorrelated in the cortex (Fig. 2a).

Odour relationships were also reshaped in the cortex compared to those in odour chemistry and boutons. UMAP embeddings of data from the tiled odour experiment (in which chain length and functional group are the main axes of chemical variation) suggested that boutons largely organize odour information along a single axis that emphasizes chain length (again, with the exception of acids) (Methods); by contrast, odour information in PCx L3 appeared largely organized in two dimensions based on functional group (Fig. 3a). Similar functional group-based reorganization was observed via hierarchical clustering (Fig. 3b). Lasso optimization confirmed that boutons and cortex differentially weight chemical features related to chain length and functional group (Extended Data Fig. 5c).

Moreover, several pairwise odour relationships were reorganized in PCx on the basis of both chain length and function group. For example, in chemical space, short-chain and long-chain odours with different functional groups were similarly cross-correlated; in boutons, correlations between short-chain aldehydes and esters were emphasized whereas those among long-chains were diminished; and in PCx L3, the opposite pattern was observed, with long-chain aldehydes and esters exhibiting stronger correlations and short-chains exhibiting weaker correlations (Fig. 3c). Chain-length-dependent cortical reshaping of odour relationships was also apparent between aldehydes and ketones.

These differences in correlation structure suggest that PCx and OB boutons differentially encode information about odour identity and odour relationships. Linear decoders based on cortical responses (particularly from PCx L3) were worse than OB-based decoders at predicting odour identity on each trial, consistent with bouton odour responses having a higher dimensionality (Fig. 4a). By contrast, cortex (particularly PCx L3) was on average better at encoding information about odour relationships (Fig. 4b); notably, however, OB was better at generalizing across short-chain ketones, aldehydes and esters whereas cortex was better at generalizing across the corresponding long-chains, consistent with the observed differences in correlation structure for these odour classes (Figs. 3c, 4c).

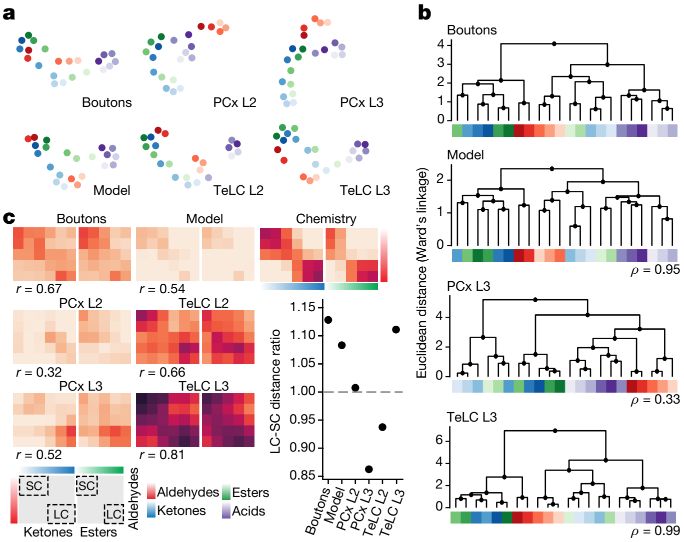


Fig. 3 | Cortical odour responses reformat odour relationships inherited from the OB. a, UMAP embeddings for all experimental conditions. Note that UMAP emphasizes relationships rather than distances, so these embeddings are similarly scaled; without such scaling the points in the model panel, for example, would be much more separated than those in the PCx L3 panel (Fig. 2a, Methods). **b**, Hierarchical clustering of neural population responses; *ρ* values indicate clustering similarity to OB boutons (Spearman correlation on cophenetic distances between boutons and the other data sets). **c**, Enlarged regions from correlation matrices in Fig. 2a depicting conserved and rearranged odour relationships between aldehydes, ketones and esters; inset: ratio of correlations between long-chain and short-chain comparisons (each dot indicates mean across odour pairs); *r* values indicate Pearson's correlation to odour chemistry. Colour code in **a** and **b** is as in **c**.

Given these differences in information content, we assessed whether bulbar or cortical odour codes more closely correspond to perceptual odour relationships by measuring the innate perceptual similarity of odour pairs through a cross-habituation assay[28] (Extended Data Fig. 8). Perceptual odour relationships better matched odour responses in PCx L3 than those in OB or PCx L2 (Fig. 4d); this closer correspondence to PCx L3 was particularly apparent for the short-chain–short-chain and long-chain–long-chain comparisons, whose pattern of neural correlation was inverted in bulb and cortex (Fig. 3c).

OB–PCx transformation requires associative network

Together, our observations suggest that the transformation between bulb and cortex reflects the combined influence of random network-type connectivity (which maintains odour relationships) and an auto-associative mechanism (which generally clusters and can selectively reshape odour relationships). To directly assess the contribution of random network-type connectivity to the observed cortical odour responses, bouton odour responses were passed through a previously established feed-forward model in which simulated PCx neurons stochastically sample from multiple glomerular inputs[19] (Methods). Consistent with previous reports, the model predicted decorrelated cortical odour representations, whose pairwise relationships were preserved relative to boutons (Fig. 2a). Although cortical responses were in part consistent with model output—as many pairwise odour relationships were preserved—the model failed to capture the strong correlation

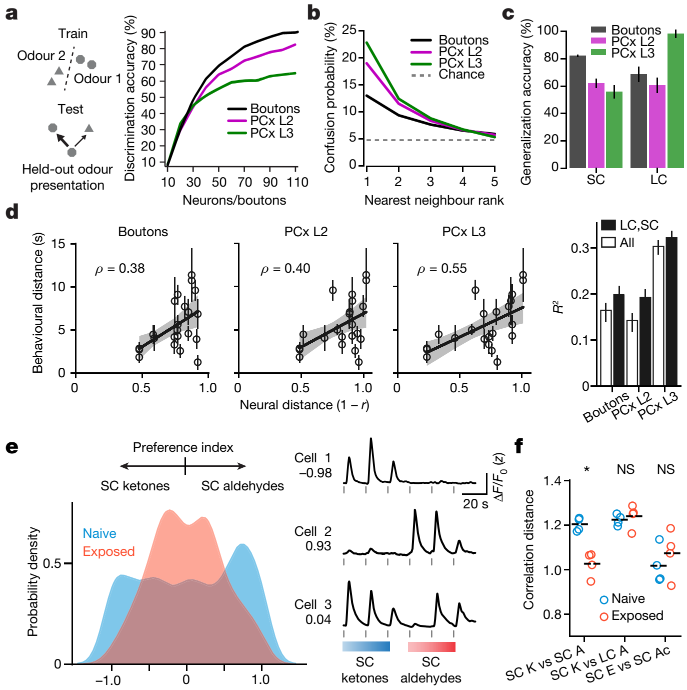


Fig. 4 | Cortical odour representations generalize across odours, are consistent with perception, and can be modified by experience. **a**, Left, schematic depicting a linear support vector machine (SVM) classifier trained to identify an odour associated with a held-out neural population response on a trial-by-trial basis. Right, decoding accuracy plotted against neural/bouton populations of different sizes. **b**, Decoding analysis to quantify odour generalization; each line represents classifier confusion between any odour and all other odours, rank ordered by the degree of confusion. **c**, Decoding accuracy of SVM classifiers predicting whether a held-out odour is a short-chain or long-chain molecule. The acid block was excluded for this analysis. Data are bootstrapped mean ± s.e.m. across held-out odours and neural/bouton ensembles. (In **a**–**c**: tiled odour set, 22 odours; number of mice, neurons/boutons for PCx L2/L3 as in Fig. 1b, d–g, for boutons, 6 mice/3,160 boutons. In **b**, **c**: 300 units, 100 bootstraps. See Methods for all decoding analyses). **d**, Left, pairwise neural and behavioural odour distances from a cross-habituation assay for the tiled odour set (Extended Data Fig. 8); *ρ* is Spearman correlation coefficient. Black line indicates regression fit (mean ± 95% confidence interval, 1,000 bootstraps). Black circles are mean ± s.e.m across mice ($n > 3$ for each comparison). Right, coefficient of determination (R^2) based on short-chain–short-chain and long-chain–long-chain or all comparisons (median ± 66th confidence interval, 1,000 bootstraps. $n = 26$ odour triplets; 122 mice across all conditions, see Methods for behavioural distance and odour identities). **e**, Left, probability density estimates of cell-wise class preference index for naive and passive odour exposure conditions, for neurons responding to at least one short-chain ketone or aldehyde (Methods). Right, example *z*-scored fluorescence (and preference index) from neurons tuned to short-chain ketones (cell 1), short-chain aldehydes (cell 2), or both (cell 3). Grey bars indicate odour onset. **f**, Pairwise odour distances in PCx L3 from odour-naive and odour-exposed mice. Passive exposure to the target mixture (short-chain ketones and aldehydes) specifically increased similarity between ketones and aldehydes, but not between control odour pairs (short-chain ketones vs long-chain aldehydes and short-chain esters vs short-chain acids (Ac)) (Extended Data Fig. 10). $^*P < 0.002$; NS (not significant) middle: $P = 0.62$; right: $P = 0.45$, two-tailed independent *t*-test. Number of mice/neurons for naive: 3/334; exposed: 3/742.

structure present in cortex or the selective rewriting of pairwise odour relationships (Figs. 2a–e, 3).

To evaluate the relative influence of auto-associative mechanisms on cortical odour representations, we used an adeno-associated virus (AAV)-based method to express tetanus toxin light chain (TeLC) within PCx neurons; this approach blocks synaptic transmission and causes PCx to behave as if it largely receives feed-forward inputs[29] (Extended Data Fig. 9). After attenuation of the associative network, the tuning of single neurons to odours broadened, response densities rose, and odour correlations increased, consistent with known network-dependent recruitment of inhibition[29] (Extended Data Fig. 9). Crucially, after TeLC infection, cortical odour relationships more closely resembled those present in odour chemistry and OB axonal boutons as assessed via correlation matrices, UMAP clustering and hierarchical clustering; for example, the cortical restructuring of short-chain and long-chain odour relationships was abolished, as was the decorrelation among acids (Figs. 2a–c, 3).

Auto-associative networks are predicted to influence correlations among odour representations to reflect the coincidence of stimuli in the world; although reward-based experiments have revealed task-dependent changes in cortical odour relationships[30,31], it has not yet been demonstrated that cortical odour correlations are sensitive to passive odour experience[1,25]. We therefore repeatedly exposed mice to a mixture of short-chain aldehydes and ketones (the PCx L3 representations of which are relatively decorrelated) (Fig. 2a) before assessing cortical responses to the tiled odour set. Mixture experience specifically increased the cortical correlation between individual aldehydes and ketones, and recruited single neuron tuning curves that reflected generalized responses to these specific odour classes (Fig. 4e, f, Extended Data Fig. 10). These observations demonstrate that cortical odour relationships can adapt to the statistics of the experienced odour environment.

Discussion

The olfactory system must synthesize information about chemical features to generate organized odour representations that support discrimination and generalization. Here we show that both OB boutons and cortex explicitly represent odour chemical relationships. The observation that many pairwise odour relationships are encoded similarly in these two brain areas is consistent with random connectivity models, which propose that PCx neurons stochastically sample glomeruli to generate a systematic population-level representation of odour chemical space[19].

However, cortex differs from the bulb in two key respects. First, PCx better clusters odour representations, enabling it to preferentially signify odour relationships. Second, cortex reconfigures information about odour relationships inherited from the bulb—the cortex does not simply pool and normalize its inputs, but instead, in a network-dependent manner, actively builds an odour space to emphasize certain odour relationships and de-emphasize others; this re-writing is sensitive to odour exposure, which can recruit new single neuron tuning properties and modify odour relationships. The olfactory system, therefore, transforms a chemical feature space into a cortical space that represents stimulus relationships through correlated activity; the structure of this space reflects information inherited from the sensory periphery, the transformation imposed by cortical circuits, and the effects of sensory experience. The cortical grouping of representations for both structurally and temporally related odours suggests a mechanism for generalization across natural odour sources, which tend to emit related odour chemicals; in principle, similar mechanisms could assign coincidentally encountered but structurally distinct odours to shared semantic categories[32–34]. Future in vivo experiments will be required to understand how the intrinsic properties of PCx neurons and the associative network, which targets both excitatory pyramidal and inhibitory neurons, collaborate to transform and organize odour representations.

In nearly all our analyses, the correlation structure of L2 odour representations was intermediate between that observed in boutons and L3, which may reflect relative differences in the prominence of bulb inputs in L2 and associational connectivity in L3[1,26]. Because the network that interconnects PCx neurons also sends centrifugal projections to the OB, it is likely that under physiological circumstances this network influences both bulb and cortical representations of odour relationships[12,29,35,36]. Although the TeLC experiment demonstrates that network activity originating in PCx is required for the bulb–cortex transformation, PCx is recurrently connected to an array of higher olfactory centres that may also have a role in shaping odour relationships[2]. Notably, neural representations in PCx L3 more closely match perception than those present in bulbar inputs, suggesting a functional hierarchy among at least some interconnected olfactory brain areas.

Relationships among cortical odour representations depend on chemical distances, such that at close distances information about chemical relationships is largely maintained, at large distances cortex decorrelates odour representations, and at the intermediate distances captured by the tiled odour set the olfactory system sculpts relational representations for odours in a manner that respects but reshapes chemical relationships. Our findings with the clustered odour set are reminiscent of previous work demonstrating that similar odour mixtures recruit overlapping ensembles of PCx neurons, although in those experiments chemical distances were not quantified[3]. Although here we take advantage of functional groups and chain lengths to systematically alter odour distances at intermediate scales, many distinct chemical features differentially contribute to odour representations in both the bulb and cortex[4]. The finding that treating odour chemicals as buckets of physiochemical features—rather than organizing information about chemistry along arbitrary dimensions—identifies structured chemical–neural–perceptual relationships is consistent with the longstanding model that the odour receptor repertoire broadly samples chemical feature space[5,10,27].

The relational information present in PCx cannot, in and of itself, assign a given odorant to its unique odour quality: the mapping observed here potentially explains why lemon characteristically smells similar to orange, but fails to explain why lemon smells like lemon. In particular, it is unclear how cortical information about odour relationships might be aligned to enable lemon odour to evoke a similar percept across individuals. We propose that relational information in PCx (and possibly other olfactory areas) is translated into invariant information about odour quality by using universal points of reference, much like a compass can be used to orient a paper map to the cardinal directions[19]. These points of reference may arise from invariant properties of specific odour receptors, or from hardwired circuits in olfactory areas such as the accessory olfactory nucleus or the cortical amygdala[15,16,37]. Alternatively, reference points could be learned from shared experience; in this model exposure to stereotyped odours (for example, amniotic fluid, mother's milk, faeces, urine or food) or common objects (such as actual lemons) could orient chemical-neural mappings along (largely) invariant axes across individuals. Further work aimed at understanding the interaction between fixed and flexible features of olfactory circuitry will be required for a full account of the relationship between chemistry, experience and perception.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2451-1.

1. Haberly, L. B. Parallel-distributed processing in olfactory cortex: new insights from morphological and physiological analysis of neuronal circuitry. *Chem. Senses* **26**, 551–576 (2001).
2. Courtiol, E. & Wilson, D. A. The olfactory mosaic: bringing an olfactory network together for odor perception. *Perception* **46**, 320–332 (2017).
3. Barnes, D. C., Hofacer, R. D., Zaman, A. R., Rennaker, R. L. & Wilson, D. A. Olfactory perceptual stability and discrimination. *Nat. Neurosci*. **11**, 1378–1380 (2008).
4. Wilson, D. A. & Sullivan, R. M. Cortical processing of odor objects. *Neuron* **72**, 506–519 (2011).
5. Haddad, R. et al. A metric for odorant comparison. *Nat. Methods* **5**, 425–429 (2008).
6. Dravnieks, A. Odor quality: semantically generated multidimensional profiles are stable. *Science* **218**, 799–801 (1982).
7. Schiffman, S. S. Physicochemical correlates of olfactory quality. *Science* **185**, 112–117 (1974).
8. Amoore, J. E. Stereochemical theory of olfaction. *Nature* **198**, 271–272 (1963).
9. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
10. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96**, 713–723 (1999).
11. Mombaerts, P. et al. Visualizing an olfactory sensory map. *Cell* **87**, 675–686 (1996).
12. Otazu, G. H., Chae, H., Davis, M. B. & Albeanu, D. F. Cortical feedback decorrelates olfactory bulb output in awake mice. *Neuron* **86**, 1461–1477 (2015).
13. Friedrich, R. W. & Wiechert, M. T. Neuronal circuits and computations: pattern decorrelation in the olfactory bulb. *FEBS Lett*. **588**, 2504–2513 (2014).
14. Chae, H. et al. Mosaic representations of odors in the input and output layers of the mouse olfactory bulb. *Nat. Neurosci*. **22**, 1306–1317 (2019).
15. Sosulski, D. L., Bloom, M. L., Cutforth, T., Axel, R. & Datta, S. R. Distinct representations of olfactory information in different cortical centres. *Nature* **472**, 213–216 (2011).
16. Miyamichi, K. et al. Cortical representations of olfactory input by trans-synaptic tracing. *Nature* **472**, 191–196 (2011).
17. Davison, I. G. & Ehlers, M. D. Neural circuit mechanisms for pattern detection and feature combination in olfactory cortex. *Neuron* **70**, 82–94 (2011).
18. Stettler, D. D. & Axel, R. Representations of odor in the piriform cortex. *Neuron* **63**, 854–864 (2009).
19. Schaffer, E. S. et al. Odor perception on the two sides of the brain: consistency despite randomness. *Neuron* **98**, 736–742.e3 (2018).
20. Roland, B., Deneux, T., Franks, K. M., Bathellier, B. & Fleischmann, A. Odor identity coding by distributed ensembles of neurons in the mouse olfactory cortex. *eLife* **6**, e26337 (2017).
21. Iurilli, G. & Datta, S. R. Population coding in an innately relevant olfactory area. *Neuron* **93**, 1180–1197 (2017).
22. Babadi, B. & Sompolinsky, H. Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
23. Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J. Neurosci*. **33**, 3844–3856 (2013).
24. Dasgupta, S., Stevens, C. F. & Navlakha, S. A neural algorithm for a fundamental computing problem. *Science* **358**, 793–796 (2017).
25. Haberly, L. B. & Bower, J. M. Olfactory cortex: model circuit for study of associative memory? *Trends Neurosci*. **12**, 258–264 (1989).
26. Bekkers, J. M. & Suzuki, N. Neurons and circuits for odor processing in the piriform cortex. *Trends Neurosci*. **36**, 429–438 (2013).
27. Saito, H., Chi, Q., Zhuang, H., Matsunami, H. & Mainland, J. D. Odor coding by a Mammalian receptor repertoire. *Sci. Signal*. **2**, ra9 (2009).
28. Cleland, T. A., Morse, A., Yue, E. L. & Linster, C. Behavioral models of odor similarity. *Behav. Neurosci*. **116**, 222–231 (2002).
29. Bolding, K. A. & Franks, K. M. Recurrent cortical circuits implement concentration-invariant odor coding. *Science* **361**, eaat6904 (2018).
30. Chapuis, J. & Wilson, D. A. Bidirectional plasticity of cortical pattern recognition and behavioral sensory acuity. *Nat. Neurosci*. **15**, 155–161 (2011).
31. Shakhawat, A. M., Harley, C. W. & Yuan, Q. Arc visualization of odor objects reveals experience-dependent ensemble sharpening, separation, and merging in anterior piriform cortex in adult rat. *J. Neurosci*. **34**, 10206–10210 (2014).
32. Sell, C. S. in *Chemistry and the Sense of Smell* **Ch. 5**, 237–296 (Wiley, 2014).
33. Gottfried, J. A., Winston, J. S. & Dolan, R. J. Dissociable codes of odor quality and odorant structure in human piriform cortex. *Neuron* **49**, 467–479 (2006).
34. Fournel, A., Ferdenzi, C., Sezille, C., Rouby, C. & Bensafi, M. Multidimensional representation of odors in the human olfactory cortex. *Hum. Brain Mapp*. **37**, 2161–2172 (2016).
35. Diodato, A. et al. Molecular signatures of neural connectivity in the olfactory cortex. *Nat. Commun*. **7**, 12238 (2016).
36. Boyd, A. M., Kato, H. K., Komiyama, T. & Isaacson, J. S. Broadcasting of cortical activity to the olfactory bulb. *Cell Rep*. **10**, 1032–1039 (2015).
37. Schoenfeld, T. A. & Macrides, F. Topographic organization of connections between the main olfactory bulb and pars externa of the anterior olfactory nucleus in the hamster. *J. Comp. Neurol*. **227**, 121–135 (1984).

Methods

Ethical compliance

All experimental procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (protocol number 04930) and were performed in compliance with the ethical regulations of Harvard University as well as the Guide for Animal Care and Use of Laboratory Animals.

Mice

Acute imaging of PCx was performed in 8–16-week-old C57/BL6J (Jackson Laboratories) male mice. Imaging of cortical neurons was performed in mice harbouring the *Vgat-ires-cre* knock-in allele (Jackson Stock no. 028862) and the *ROSA26-LSL-TdTomato cre* reporter allele (Jackson Stock no. 007914); imaging of boutons was performed in mice containing the *Tbx21-cre* allele (Jackson Stock no. 024507, gift from C. Dulac). TeLC-dependent elimination of cortical excitatory transmission and subsequent imaging was performed in Emx1-IRES-Cre mice (Jackson Stock no. 005628). Male mice were group-housed before viral delivery of GCaMP6s and housed singly for 1–3 weeks after injection.

Viral constructs

To generate pAAV-hSyn-FLEX-TeLC-P2A-dTom, pAAV-hSyn-FLEX-TeLC-P2A-EYFP (a gift from Bernardo Sabatini) was digested with AscI and NheI to remove enhanced yellow fluorescent protein (eYFP). A gene fragment (synthesized by IDT) containing dTomato with the SV40 nuclear localization signal was cloned into the TeLC backbone via isothermal assembly (NEB HIFI E2621). AAVDJ-hSyn-FLEX-TeLC-P2A-dTom virus was produced by Vigene Biosciences, with a titre of 1.5×10^{13} genome copies per ml. AAV PHP.eB hSynapsin1-FLEx-axon-GCaMP6s virus was produced as previously described[38]. In brief, HEK293T cells were co-transfected with pAAV-hSynapsin1-FLEx-axon-GCaMP6s (a gift from L. Tian, Addgene plasmid 112010), PHP.eB rep-cap (a gift from the Viviana Gradinaru Lab/Clover centre), and adenovirus helper plasmids. After 5 days, viral particles were collected and then purified via iodixanol gradient ultracentrifugation. The virus was titred via quantitative PCR (qPCR); the titre of all batches was between 2×10^{13}–3×10^{13} viral genomes per ml.

Stereotaxic viral delivery

Vgat-ires-cre; ROSA26-LSL-TdTomato male mice were injected with an AAV expressing the genetically encoded activity indicator GCaMP6s (AAV1.CAG.GCaMP6s.WPRE.SV40, Penn Vector Core). Injections were targeted to posterior PCx using Allen Brain Atlas coordinates: medial–lateral (ML): −4.2, anterior–posterior (AP): −1.09, dorsal–ventral (DV): −4.25, from the dura. FLEX-TeLC-P2A-dTom was targeted to anterior and posterior PCx of both hemispheres of Emx1-IRES-Cre mice at AP: 0.39 and −1.0; ML: −3.51 and −4.2; DV: −4.4 and −4.1. To uniformly infect olfactory bulb projection neurons, AAV PHP.eB FLEX-axon-GCaMP was delivered intravenously via retro-orbital injections in Tbx21-Cre transgenic mice (Jackson Stock no. 024507), which express Cre recombinase in OB projection neurons.

Full-titre viruses (500–1,000 nl) were delivered to cortex at 1 nl $\mathrm{s^{-1}}$ using a Nanoject II dispensing pump (Drummond Scientific). GCaMP6s-injected mice were imaged 1–3 weeks after delivery. In TeLC experiments, GCaMP6s was injected at ML: −4.2, AP: −1.09, DV: −4.25, from the dura, 2–3 weeks after TeLC delivery. Mice were imaged 3–5 weeks after TeLC delivery. Imaging of OB projection neuron axon terminals in L1a of PCx was performed 3–5 weeks after retro-orbital injection. To assess the influence of passive odour exposure on cortical representations (see 'Passive odour exposure'), odour exposure to mixtures was initiated 1–2 weeks after viral delivery; as with imaging in odour-naive mice, imaging in odour-exposed mice was performed 3–4 weeks after injection.

Uniform infection of L2 and L3 cortical neurons by AAV-TeLC across the rostro-caudal extent of PCx was confirmed histologically. To validate TeLC dependent inhibition of cortical excitatory synaptic transmission, odour-evoked single-unit activity was compared between control and infected mice as previously described[29].

Surgical approach and craniotomy

We developed a surgical preparation compatible with PCx imaging during wakefulness and semi-paralysis; this preparation is similar to those used in the past to explore in vivo neural responses without the use of general anaesthesia (but with effective analgesia) during the experiment[39–41]. Before exposing PCx, mice were anaesthetized with isoflurane, and head-fixed with dental cement to a rotating headpost. The PCx was then accessed from the ventral surface of the mouse skull through surgical resection of the zygoma, the mandible, and associated musculature and fascia. A 2 mm craniotomy overlying the PCx was made using a dental drill and secured with a custom-shaped cranial window.

To ensure that the mouse was free of pain and discomfort during wakefulness and semi-paralysis, full hemifacial analgesia was provided by performing a complete trigeminal nerve block. This procedure is designed to abolish sensation around the surgical exposure, as well as the ipsilateral oro-facial region encompassing the entire dorsoventral extent of the head and extending from the nostril to the neck. The junction of the four branches of the trigeminal ganglion was readily identified at the external pterygoid ridge, which was rendered accessible when the mandible was removed. A 1–5 µl, 0.2–1.0 mg $\mathrm{kg^{-1}}$ dose of bupivacaine was injected directly into the stalk of the trigeminal nerve bundle using a calibrated micropipette mounted on a micro-manipulator. By infusing bupivacaine solution proximally to the trigeminal ganglion, distribution along all trigeminal branches, including the mandibular, ophthalmic, and infraorbital branches, was ensured. To verify that the block infiltrated the entire nerve bundle, each injection was supplemented with the fluorescent lipophilic contrast agent DiI used to identify myelinated nerve fibres, owing to its lipophilic, infiltrating nature[42]. By including DiI in the block solution and monitoring its diffusion through the nerve adjacent to the injection site, proper micropipette placement directly inside the nerve bundle was confirmed. Successful DiI injections were characterized by uniform distribution of dye through the trigeminal bundle proximal to the injection site. In cases of insufficient labelling of the nerve bundle, several injections were administered until the entire nerve was labelled by visual inspection. This procedure was extensively evaluated through measurements of heart rate (which revealed no signs of distress)[43] (Extended Data Fig. 2) and by systematically probing the depth of analgesia through the use of needle pricks along the entire dorso-ventral and rostro-caudal portion of the head ipsilateral to the injection site.

After completion of the surgical exposure, induction of analgesia, installation of paralytic infusion and retro-nasal sniffing lines, as well as placement of the EEG electrode (see 'Semi-paralysis', 'Artificial sniffing' and 'EEG for assessing brain state' sections), isoflurane anaesthesia was discontinued and mice were transferred to the imaging set-up equipped with custom-built sniff generator, oxygen respirator, as well as a peristaltic pump for paralytic infusion.

Semi-paralysis

Mice were provided with a continuous infusion of a low dose of the muscle relaxant pancuronium bromide into the jugular vein during the imaging phase of the experiment[40]. After calibration, the final dose and infusion speeds were chosen to be 0.024 ng $\mathrm{kg^{-1}}$ per 10 min. At this dose, mice experience a loss of righting reflex, but maintain diaphragmatic contractions and toe-pinch reflexes. Because this dose was chosen as to minimize paralysis (which is not required for analgesia), if movement was observed in the experiment intermittent pushes of pancuronium were provided to ensure motion-free imaging. We refer to this preparation as being in the condition of 'wakefulness' (see 'EEG for assessing brain state') as opposed to 'awake' given that the mice are incapable of gross movements and cannot actively sample odours because of the ventilator (see 'Artificial sniffing').

Artificial sniffing

To control for potential differences in odour coding due to changes in odour sampling, the mouse's sniffing rhythm was replaced with an 8-Hz fixed inspiration/expiration cycle synchronized in time to odour presentation; this cycle rate mimics known sniffing rates during active odour sampling[44,45]. We adopted a previously developed method in which a cannula was placed into the nasopharynx via the trachea and subsequently attached to a solenoid valve which draws air bidirectionally across the nasal epithelium[46]. The tube was secured to the trachea with a pair of nylon sutures and doused with silicone elastomer for further stability[47]. The distal portion of the tube was then coupled to a computer driven solenoid valve and a vacuum line, providing 50-ms pulses of suction every 75 ms at a flow rate of 100 ml min^{-1}.

EEG for assessing brain state

Anaesthesia is thought to induce a brain-state similar to slow wave sleep that is characterized by large-amplitude fluctuations in the 0.5–4.0 Hz range and the absence of high-frequency activity from 40–100 Hz, which is typically present during wakefulness or behavioural engagement. Because power in the slow and fast frequency bands of the EEG is anti-correlated across these brain states, their ratio has been traditionally used to assign an absolute value to the arousal state of the mouse[48,49]. To compare changes in brain state between wakefulness and anaesthesia and associated changes in odour representation, after completion of awake imaging, some mice were subsequently anaesthetized with a ketamine dose of 50 mg kg^{-1} and 0.5 mg kg^{-1} of medetomidine delivered together intraperitoneally, and the imaging session was immediately repeated. For all experiments, EEG activity was recorded using a silver wire inserted into the dorsal anterior PCx via a 0.5-mm craniotomy. A grounding wire was placed into the contralateral cerebellum. This signal was amplified using an AM Systems 1800 amplifier and digitized with a National Instruments PXie-6341 acquisition card. Signals were detrended and bandpassed (0.5–500 Hz) before computing the EEG power ratio.

Odour space design

A major goal of this study was to rationally design odour sets such that chemical similarities and differences between odorants in each odour panel could be explicitly titrated. As described previously, to describe odour space we took advantage of 2,584 molecules commonly used in the flavours and perfume industries from http://www.thegoodscentscompany.com/[27]. A large fraction of these molecules are odorous, in that they are less than 300 Da in size and sufficiently hydrophilic and volatile to readily access the environment of the nasal epithelium. This collection contains structurally diverse molecules that vary in carbon chain length, weight, polarizability, hydrophobicity, cyclicity, branching, constituent functional groups, and other chemical attributes. To characterize the physiochemical features of the odours within this large odour collection, we took advantage of a database of 3,705 statistical metrics designed to quantify different molecular physiochemical properties (Dragon, KODE Inc.), including those related to molecular weight, volume, ionization potential, and so on. Using this descriptor database, each of the 2,584 collected molecules were represented by a vector containing 3,705 values (in which each value is a quantitative description of a specific physiochemical feature), thereby constituting an odour chemical space where odour similarity can be expressed as the Euclidean, cosine or correlation distance between any two molecules[5,50]. Of these features, 2,522 are quantified in the Dragon database as continuous variables (for example, molecular weight) and 985 were binarized or categorical (for example, the presence or absence of a N atom). The collection of these inter-odour distances (which holistically capture the quantified physiochemical differences between each odour pair) can be converted into correlation distance matrices (see 'Chemical and activity distance').

Odour selection

Three ('global', 'clustered' and 'tiled') distinct odour sets were identified, each consisting of 22 odorants. The global odour set contained structurally diverse molecules that span the entire odour space. The clustered odour set consisted of 6 groups of 3–4 molecules, such that all the odours within each group share a chemical functional group (as well as other common chemical features); these groups were designed such that the odours that belong to each group were maximally separated from the odours belonging to all other groups. The tiled odour set included closely related aliphatic molecules that systematically varied along two dimensions; the first was the number of carbon atoms in the chain, and the second was the particular functional group attached to the carbon chain (that is, aldehydes, esters, ketones and acids, all of which are related to each other). Odour selection for the first two odour sets was performed with stochastic optimization (see 'Simulated annealing') to prevent human-induced biases in odour set design. The cost function for the global odour set was designed to maximize separation between all 22 odours (by maximizing the minimum pairwise distance among selected odours). For the clustered odour set, within-group similarity and between-group dissimilarity was maximized by using the silhouette coefficient as the cost function.

Odour delivery

A 23-valve olfactometer that can deliver up to 22 odorants was used to present odours (Island Motion). The 23rd valve was used to deliver a blank stimulus (no odour) between odour presentations. Custom Arduino software was used to control valve opening and closing, thereby enabling switching between odour vials and the blank vial. This software also controlled the output of two mass flow controllers (MFC). The first MFC delivered a constant carrier flow at 0.8 l min^{-1} of purified air into a common channel; the second MFC supplied a constant flow at 0.2 l min^{-1} of clean air that was injected into an odour vial (see below) and then merged with the carrier flow 1 inch (2.54 cm) in front of the mouse's nose. A larger exhaust fan drew air from the imaging cage that enclosed the rig to prevent cross-contamination.

Monomolecular odours were diluted in di-propylene glycol (DPG) according to individual vapour pressures obtained from www.thegoodscentscompany.com, to give a nominal concentration of 500 ppm. This vapour-phase concentration was further diluted 1:5 by the carrier airflow to yield 100 ppm at the exit port. Odour presentations lasted for two seconds and were interleaved by 30 s of blank (DPG) delivery. The order of presentation of odours was pseudo-randomized for each experiment, such that on any given trial, odours were presented once in no predictable order. Each odour was presented 7–10 times in each experiment.

Two-photon calcium imaging

High-speed volumetric imaging was performed using a 16-kHz resonant galvo-regular galvo pair (Cambridge Technologies) housed in a custom-designed microscopy rig equipped with 2-inch optics. Acquisition was performed with a large working distance Nikon 16× objective (N16XLWD-PF, 0.8 NA, 3 mm WD) mounted on a high-speed piezo actuator (nPoint 400). A Chameleon laser (Coherent) tuned to 930 nm delivered 50–120 mW of excitation power at the front end of the objective. Emitted fluorescence was detected using Hamamatsu H10770PA-40 PMTs. Scanimage 5 was used for hardware control and data acquisition.

For imaging neuronal cell bodies, acquisition volumes spanned 210 µm in the *Z* axis across PCx L2 and L3. Volumes were split into 6 optical slices each spanning 35 µm of cortex. Volumes were positioned such that two slices resided in L2 and four slices resided in L3. This allowed us to monitor similarly sized populations of neurons in L2 and L3 given the approximately threefold lower cell density of L3 in posterior PCx[51]. We typically discarded a single optical slice that spanned the boundary between layers to avoid any cross-contamination between layers. For axonal imaging, a single plane was acquired in PCx L1a at 60 Hz and subsequently downsampled by averaging to match the neural acquisition rate.

Note that our imaging fields are in the most anterior portion of posterior PCx. The degree of associational connectivity is known to systematically vary across the anterior-posterior axis of the PCx, with the least associational and most feed-forward connectivity anteriorly[52,53]. We chose to image in the 'middle' of the PCx both because of anatomical constraints in our imaging field in the two-photon configuration, and to ensure the representations we probed would include both feed-forward and associational connectivity. We would expect that if we imaged more anteriorly, we would observe representations that were progressively more 'bulb'-like (given the relative predominance of inputs), and conversely that posterior PCx would deviate more strongly from the bulb (given the relative predominance of associational connectivity); because of surgical constraints, addressing this possibility will require the future development of alternative means of accessing anterior and posterior PCx both pre- and post-synaptically.

Data inclusion criteria

For experiments involving the global, clustered and tiled odour sets in odour-naive mice, data were analysed from three mice per odour set. For the passive odour exposure experiment, data were analysed from three mice. For bouton imaging, data was analysed from six mice. All mice that satisfied the following pre-determined criteria were included in the study: imaging volumes spanning both piriform cortical L2 and L3 (L1a for boutons) could be imaged continuously for the duration of the experiment; in each cortical layer, at least 150 GCaMP6s-labelled neurons could be identified (500 axonal boutons for axonal imaging); odour-evoked activity persisted over the course of the entire imaging session; and field-of-view drift and motion artefacts could be fully corrected with post hoc image registration. Given the nature of this population imaging study, study sample size was not pre-determined, the experiments were not randomized, and the investigators were not blinded to study conditions.

Signal extraction

Detection of regions of interest (ROIs), segmentation, and extraction of fluorescence signal was performed using the Suite2p software[54]. This package implements image registration, neuropil fluorescence correction and fluorescence source detection from spatially overlapping ROIs. To accommodate differences in ROI size between axonal boutons and somata, the expected ROI size paramater was set to 5 μm for axonal boutons and 12 μm for somata.

auROC-based detection of odour responses

Analysis was only performed on neurons that responded, in a statistically significant manner, to at least one odorant. To identify such neurons, we computed the area-under-the-receiver-operator-curve (auROC) statistic for each cell-odour pair. The auROC metric represents the probability that a neuron's response, chosen at random from all presentations of the same odour, will be ranked higher than a randomly chosen sham response obtained using baseline activity. A value of 0.5 indicates no difference between the activity of a neuron during baseline and odour presentation. A value of 1 indicates a perfectly distinguishable excitatory response, whereas a value of 0 indicates a perfectly distinguishable suppressed response. For a single neuron and all presentations of a single odorant, the classifier was provided with the mean fluorescence obtained from 2-s time windows immediately flanking odour onset. A null distribution of auROC values for each cell–odour pair was constructed by randomly permuting the identity of the odour and baseline periods on each presentation. This procedure was repeated 1,000 times. The actual auROC value was deemed significant if it resided outside the 1–99th percentile of the null distribution. Neurons that did not display a significant response to any odours, according to auROC analysis, were excluded from all subsequent analysis. Of those neurons imaged, the fraction of retained neurons (and the absolute number of neurons in each data set) were: global L2 = 854 neurons, 82%; global L3 = 616 neurons, 89%; clustered L2 = 867 neurons, 87%; clustered L3 = 488 neurons, 85%; tiled L2 = 427 neurons, 59%; tiled L2 = 334 neurons, 52%; TeLC tiled L2 = 435 neurons, 51%; TeLC tiled L3 = 590 neurons, 68%; boutons tiled = 3,160 boutons, 68%. Note that the number of neurons deemed responsive by auROC analysis is proportional to the extent to which each odour set captured chemical diversity, with the greatest number of responsive neurons observed in the global odour set, and the fewest observed in the tiled odour set. This distribution of responsive neurons (between 51% and 89%, depending on the chemical diversity in each odour set) is consistent with previous work characterizing response densities and tuning breadths in PCx.

Gaussian mixture model for response type clustering

Clustering of cell–odour response types was performed for visualization purposes only. Trial-averaged response time courses spanning the odour presentation period were dimensionally reduced by principal component analysis (PCA) to capture 90% of the variance in the data and served as the input to a Gaussian-mixture model, with the optimal number of clusters was assessed using the Bayesian information criterion.

Lifetime and population sparseness

Lifetime sparseness is a metric reflecting the tuning breadths of individual neurons, with neurons specifically tuned to small numbers of stimuli exhibiting a lifetime sparseness of close to 1; population sparseness is a metric reflecting the density of responses among a population of neurons to a set of stimuli, with less dense responses (that is, fewer neurons or boutons responding to a stimulus set) exhibiting a population sparseness of close to 1. To determine the odour-selectivity of a neuron, the lifetime sparseness metric was computed as previously described[55]:

$$\text{lifetime sparseness} = \frac{1 - \left(\sum_j^N \frac{r_j}{N}\right)^2 \Big/ \left(\sum_j^N \frac{r_j^2}{N}\right)}{1 - \frac{1}{N}}$$

where r_j is the positive odour-evoked change in fluorescence to an odour j relative to baseline and averaged over multiple odour presentations, and N is the number of odours (22 in all odour sets). Inhibitory responses were zeroed (for this analysis only). Lifetime sparseness reflects the kurtosis of the tuning profile of a neuron and ranges from 0 to 1. Highly peaked, narrow tuning profiles yield values close to 1 and represent neurons that respond strongly and selectively to few odours. Values close to 0 indicate equal responsiveness to a large fraction of the odour set. Population sparseness for each odour was calculated using the same formula used for lifetime sparseness, but in this case, j indexes a neuron instead of an odour.

Signal and ensemble correlations

The extent to which any two neurons have similar odour preferences can be assessed by computing the Pearson's product moment correlation between their trial-averaged odour response profiles (tuning curves). This is typically referred to as a 'signal' correlation. The tuning curve of each neuron was represented as a vector containing N elements, where N is the number of odours in the stimulus set. Each entry in this vector corresponds to the odour-evoked change in fluorescence relative to baseline and averaged over all presentations. For each neuron, responses across odours were *z*-scored. Populations were defined as all neurons that responded to at least one odour according to the auROC analysis.

We also wished to compute the similarity in odour responses exhibited by different odour-evoked ensembles of PCx neurons or boutons. We refer to this here as 'ensemble' correlation. In these analyses, we computed the Pearson's correlation between the population responses to every pair of odours in our panel. The population response of each odour was represented as a vector containing N elements, where N is the number of neurons/boutons in the population. Each entry in this vector represents the trial-averaged response of each single neuron/bouton to the corresponding odour. For each neuron/bouton, responses across odours were z-scored. Populations were defined as all neurons/boutons that responded to at least one odour according to the auROC analysis.

Chemical and activity distance

Pairwise odour distances (1 − Pearson's r) in neural activity space were computed between odour vector pairs where each matched vector entry corresponded to a single neuron's trial-averaged response to the corresponding odour. This procedure was applied to neural populations from individual mice or to pseudo-populations of neurons built by pooling all responsive neurons from all mice for each layer and odour set. In chemical space, correlation distances between odour pairs were computed identically, except each vector entry (matched across odours) represented the odour-specific value assigned by a physiochemical descriptor. For presentation purposes, distance matrices were sorted using hierarchical clustering. For the global odour set, all odours were sorted collectively. For the clustered odour set, odours were sorted within each functional class first followed by sorting on functional classes. For the tiled odour set, functional group classes were sorted, but odours within each class were ordered according to increasing chain length. Row and column ordering of all activity and chemistry distance matrices is preserved across figures. Note that for all correlation analyses, both inhibitory and excitatory responses were included.

UMAP embedding

For visualizing odour relationships in neural data, population responses were embedded in two dimensions using UMAP[56]. Selection of optimal embedding settings was accomplished by minimizing the mean-squared error between correlation distance matrices built from data projected on the UMAP dimensions and those corresponding to the input data. Simulated cortical responses from the feed-forward models were processed in a similar manner. Because UMAP imposes an arbitrary rotation on projected data, each embedding was aligned to a reference using the orthogonal Procrustes transformation. For embeddings of pseudopopulation data for the tiled odour set across boutons as well as neural and simulated cortical data, bouton data served as the reference. For aligning embeddings obtained from individual mice, the orthogonal Procrustes transform (rotation and reflection only) was performed iteratively across mice in a pairwise manner[57]. Note that these embeddings are meant to visualize odour relationships (and are complemented by quantitative metrics); pairwise relationships cannot be changed by any of the rotations used herein to align embeddings to each other.

Distance-based nearest neighbour decoding and classification

To test whether PCx odour relationships are invariant across individuals, we asked whether we could identify any given odour from one mouse based upon the pairwise odour relationships observed in other mice. To decode odour identity based on odour relationships, nearest neighbour classifiers were trained on odour distances from two mice and tested on odour distances obtained from a held-out mouse, such that each mouse was tested once. For each such classifier, 100 bootstrap iterations were performed. For each iteration, correlation distance matrices were constructed using a randomly sampled neural ensemble containing 50 neurons. In each condition, each distance matrix represented all pairwise correlations between trial-averaged population responses (see 'Chemical and activity distance'), such that any given training or testing odour was represented as a vector containing 21 pairwise distances. For a single run of the classifier, reported accuracy represents the fraction of odours that were correctly identified.

Distance covariance analysis

Distance covariance analysis (DCA) belongs to a set of statistical methods that seek to identify shared dimensions of variability between two different data sets. DCA—an extension of canonical correlation analysis—was developed for identifying related dimensions of activity across two or more populations of neurons[58]. To measure the similarity of cortical odour relationships between individual mice, DCA was performed on response data (neurons by odour-trial) from all individual mice exposed to the same set of odours (code supplied with the reference). The output consisted of a set of orthogonal dimensions (one set per mouse) and associated DCA statistics ranked from highest to lowest contribution to common activity between individuals. Each dimension was evaluated for significance by permutation testing. The null distribution of DCA statistics was constructed by shuffling the sequence of odour responses across all neurons in each mouse to destroy between but not within-mouse relationships. A dimension was deemed significant if its associated statistic was higher than the 95th percentile of the null distribution built from 100 permutations. Three to six dimensions were typically retained. Because DCA is not deterministic, this procedure was subjected to 100 independent restarts. The reported results correspond to the best modelling run. The fraction of an individual's neural variance that could be explained by the shared embedding was determined by calculating the total accounted variance after regressing each neuron's activity on the set of DCA dimensions.

Silhouette coefficient

The degree of clustering in odour correlation distance matrices was evaluated using k-means clustering and the silhouette coefficient. Correlation distances were computed between trial-averaged population responses to all 22 odours in the tiled odour set. Correlation distance matrices containing all pairwise odour distances were projected onto 21 principal components and subjected to k-means clustering. For this set of labelled data, the silhouette coefficient assigns a single value ranging from −1 (overlapping diffuse clusters) to 1 (compact, well-separated clusters) that represents the average silhouette score computed on an odour-by-odour basis: for an odour i, its score, S_i, is defined as $(b_i - w_i)/\max(w_i, b_i)$, where w_i is the average Euclidean distance between odour i and all other odours with the same class label, and b_i is the average distance between odour i and all odours in the next nearest class. Qualitatively similar results between experimental conditions were obtained by running k-means and cluster evaluation directly on full population response data or by obtaining k-means labels from full population data and computing the silhouette coefficient using these labels on PCA embeddings of correlation distance matrices.

Effective dimensionality

Effective dimensionality (ED) of a population of neurons, a quantity reflecting the number of principal components required to capture the odour-evoked neural variance, was defined as previously described[59,60]. In brief, for each experimental condition and for model cortical activity, ED was quantified from trial-averaged population responses (neural data only) after mean-centring units across odours. ED reflecting variance in similar ensemble sizes were calculated as averages across randomly chosen, 300-unit ensembles (100 bootstraps).

Hierarchical clustering

Dendrograms depicting the reconfiguration of odour relationships across boutons, PCx, TeLC, and the feed-forward random connectivity model PCx were constructed directly from correlation distance matrices associated with each experiment. First, each correlation matrix was projected, using PCA, onto K dimensions, where K is the number of dimensions required to explain 95% of the variance in correlation distances. The resulting embedding expresses the contributions of each odour to the prominent similarity or dissimilarity modes in the original correlation distance matrix. Dendrograms were built by hierarchically clustering this data using Euclidean distance and Ward's linkage. Clustering similarity between dendrogram pairs was assessed using the Cophenetic correlation coefficient after topologically aligning each pair.

Feed-forward connectivity model

To assess whether the observed cortical odour responses are expected under a random feed-forward model we simulated the OB–PCx network using a previously established model[19]. In our implementation of this model (which hews as closely as possible to the published model), the OB and PCx layer contained 1,000 and 100,000 units, respectively, feed-forward connections were assigned randomly and can be either excitatory or inhibitory, and each cortical neuron was 'innervated' by a random 20% of excitatory OB units and a random 40% of inhibitory units. Excitatory and inhibitory connection weights were set to 1 and −0.5 respectively, providing each PCx unit with balanced excitatory/inhibitory innervation. Odour-evoked activity in the OB layer was simulated from a log-normal multivariate distribution defined by population-mean response amplitudes and covariance obtained from bouton activity. Model PCx neurons linearly summed their inputs and are zero-rectified. The odour-average response fraction of model PCx units was adjusted to 8% to match the fraction of excitatory responses observed in PCx (detected by auROC analysis) on average across all odours and subjects in our experimental data.

Decoding analysis

Linear SVM classifiers were trained to predict either odour identity or odour class (based on chain-length) in the tiled odour set on the basis of odour-evoked population activity.

For discrimination of odour identity (Fig. 4a), all neurons or boutons that responded to at least one odour in the tiled odour set (according to auROC analysis) were pooled to build three pseudo-populations of neurons or boutons (boutons, L2, L3). *Z*-scored responses of a population of up to *n* randomly selected neurons or boutons (the maximum common number of neurons recorded across the layers) were then considered, given *t* presentations of *j* odours as a matrix X with *n* rows (neurons/boutons) and $t \times j$ columns (trials/instances × odours/classes). Each column of this matrix was thus a vector of *n* responses, one for each neuron/bouton in response to a given odour in each trial. Each decoding session started with a split of the matrix X into a training and test set: the training set included $0.9 \times t$ randomly chosen trials for each class and the test set comprised the $0.1 \times t$ held out trials for each class (that is, a standard 9:1 training:testing split). This procedure, which is instantiated as part of the standard LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), allows us to use a binary classification algorithm (such as an SVM) to compare multiple classes. In any given experiment, the train-test procedure was iterated 100 times (with training and test data randomly chosen on each iteration) to cross-validate classifier performance. For differently sized subpopulations of neurons or boutons, a randomly selected subset of neurons or boutons was used for each cross-validation cycle, and at the end of this procedure the outcomes of each individual iteration were averaged to generate a measure of classification accuracy across all restarts; this is the overall measure that is reported in the main text (Fig. 4a). The hyperplanes for each classifier were determined using the LIBSVM library with a linear kernel, the C-SVC algorithm, and cost *c*. Cost *c* is the only free parameter for a linear kernel, and it was found by a grid search on an initial data set including 50 randomly chosen neurons/boutons from each dataset in order to maximize the accuracy of the decoder classification.

For Fig. 4b, classifiers were trained on 21 out of 22 odours in the tiled odour set, with all trials associated with any training odour assigned to 1 out of 21 classes. SVM class predictions for each held-out odour were converted to confusion probabilities (the probability that any given held-out odour is associated with any other of 21 odours) using the Python scikit-learn library[61] implementation of Platt scaling[62]. Class probabilities for each tested odour were rank-ordered before averaging across all odours.

For Fig. 4c, classifiers were trained to predict the class (either short-chain or long-chain) of a held-out odour after training on a single short-chain–long-chain odour pair. All odour trials were presented on each train-test iteration and accuracy was determined as the fraction of correctly labelled held-out trials. For each randomly chosen subpopulation of neurons, on each of 100 restarts, training and testing was performed on all possible short-chain–long-chain odour pairs and associated held-out odours. Cross-validated generalization accuracy corresponds to the average performance across all restarts and folds of the data. The short-chain class contained aldehydes: propanal, butanal and pentanal; ketones: propanone, butanone and pentanone; esters: ethyl and butyl acetate. The long-chain class contained aldehydes: heptanal and octanal; ketones: hexanone, heptanone and octanone; esters: pentyl and hexyl acetates. The acid block is excluded for this analysis.

Lasso optimization

For finding small, optimal combinations of Dragon descriptors that predict neural odour relationships, an L1-regularized optimization routine was designed to maximize the correlation between matched odour-pair distances across chemical and neural spaces. During each step of optimization ('L-BFGS-B' gradient descent), descriptor weights (0 bounded) were modified and chemical distances were recomputed. The optimization objective sought to minimize the residual sum of squares between the modified descriptor distances and corresponding neural odour distances. The Lasso component set the sparseness of the final solution and was selected for each model by cross-validation. Models were trained and validated on individual odour sets containing 22 molecules with fivefold cross-validation (random splits) such that on any split of the data, $17 \times (17-1)/2$ odour pairs made up the training set and $5 \times (22-5)$ odour pairs comprised the validation set. For assessing generalization to other odour sets, models were retrained with all 22 odours before testing. Because the global and clustered odour sets share odours, overlapping odours were removed from training when cross-applying to the held-out odour set. For within-odour set cross-validation and cross-application to the tiled odour set all odours were included. For Extended Data Fig. 5c, we sought to determine the relative contribution of the full set of descriptors belonging to the 'molecular properties' block of the Dragon database to bouton/cortical relationships; for this analysis, optimization was performed without imposing L1-regularization.

Simulated annealing

Stimulated annealing was used for odour set design, as well as for predictive modelling of neural odour relationships. Simulated annealing is a well-validated Monte Carlo sampling variant designed for stochastic optimization[63]. Simulated annealing optimization works by slowly decreasing a pre-specified cost function over the course of sampling, thereby enabling good initial coverage of solution space and progressive convergence on a global optimum. All simulated annealing routines were implemented using the open-source Python package simanneal available at https://pypi.python.org/pypi/simanneal. For odour set selection, the number of features was reduced, via PCA, such that the transformed odour space accounted for 95% of the original variance. Optimization for the global and clustered odour sets (see 'Odour selection') was carried out using Euclidean distance in this reduced space.

In addition, our findings using Lasso optimization were verified using simulated annealing; here, the simulated annealing objective was designed to identify small sets of Dragon physiochemical features describing a set of molecules such that molecular distances in chemical space were maximally correlated with corresponding odour distances in neural space. Qualitatively similar findings using simulated annealing were observed as reported in the manuscript using Lasso optimization (data not shown). By design, the chemical variation in the tiled odour set, queried here, is focused on chain length and functional group; while the observation that bulb and cortex differentially encode information about these chemical features clearly demonstrates a representational transformation, it does not imply that chain length or functional group per se are privileged in either of these brain regions relative to the multitude of chemical features not captured by the tiled odour set.

Cross-habituation for assessing perceptual odour similarity

C57 males (5–6 weeks old) obtained from Jackson laboratories were housed on a reverse light schedule for 48 h before beginning behavioural experiments. Established procedures for assessing odour similarity were slightly modified[64]. Twenty-six pairwise comparisons were obtained, with twelve odorants serving as the first odorants in each comparison: butanal versus pentanal or propyl acetate; butanoic acid versus heptanoic acid or butanone; butanone versus butanoic acid or propyl acetate or butanal; heptanal versus octanal, or pentyl acetate; heptanone versus pentyl acetate or heptanal or hexyl acetate or octanal; hexyl acetate versus pentyl acetate or octanone or octanal or octanoic acid; octanal versus butanal or heptanal or octanone; octanoic acid versus octanal; octanone versus octanoic acid; pentyl acetate versus hexyl acetate; propanoic acid versus butanoic acid; propyl acetate versus butanone or butanal or butanoic acid.

Short-chain pairwise odour comparisons, different functional groups (asterisks indicate the same pair was presented at different positions in the triplet): butanone versus butanal; butanone versus butanoic acid*; propyl acetate versus butanone*; propyl acetate versus butanal*; propyl acetate versus butanoic acid.

Long-chain pairwise odour comparisons, different functional groups (asterisks indicate the same pair was presented at different positions in the triplet): heptanal versus pentyl acetate; octanal versus octanone; heptanone versus pentyl acetate; heptanone versus heptanal; heptanone versus hexyl acetate; heptanone versus octanal; octanone versus octanoic acid; pentyl acetate versus hexyl acetate*; hexyl acetate veresus octanone; hexyl acetate versus octanal; hexyl acetate versus octanoic acid; octanoic acid versus octanal.

Remaining pairs: heptanal versus octanal; octanal versus butanal; butanal versus pentanal; propionic acid versus butanoic acid; butanoic acid veresus heptanoic acid.

Because three odours were presented to each mouse, two adjacent odour pairs were included in analysis from each mouse. Presentation order effects were considered by swapping the order of any given triplet in different experiments. As indicated above, for some comparisons the same pair was presented at different positions in the triplet. Investigation time was scored manually, using video footage obtained during each experiment. Scoring was done blinded to experimental conditions, and with no knowledge of odour identity. Odour investigation was defined as periods of orienting to the odour source on the half of the cage containing the odour source as well as by stereotyped bouts of sniffing and associated head-bobbing.

Perceptual similarity (behavioural distance) between two odours was defined as the difference in time spent investigating the first odour in a pair during its last presentation and the investigation time associated with the first presentation of the subsequent odour (Extended Data Fig. 8). In Fig. 4d, linear regression was used to relate behavioural to neural distance. Because behavioural distance increases monotonically but not necessarily linearly with neural distance, we also used Spearman ρ, which measures correlation based on ranks and is less restrictive than linear regression. Three to twelve mice were used for each pairwise comparison (eight mice on average per triplet experiment). Each mouse was used for a single set of odour comparisons.

Passive odour exposure

C57 males (5–6 weeks old) were housed on a reverse light schedule for 48 h before behavioural training. Group-housed mice were subjected to daily odour exposures for a period of two weeks. On each training session (30 min; 3 times per day, 14 consecutive days) mice were simultaneously presented with two short-chain aldehydes (propanal, butanal) and two short-chain ketones (propanone, butanone) for 1 min separated by a 5-min inter-stimulus-interval. Odour delivery was designed to closely approximate odour presentation during cortical imaging. In brief, odours were delivered to the home cage using a custom-built olfactometer consisting of an activated-carbon purification unit, master air flow controllers, and a valve bank coupled to four odour vials and one blank vial. Flow rates for carrier and odour lines were set to 0.8 and 0.2 l min^{-1} respectively. Monomolecular odorants were diluted in DPG according to individual vapour pressures to yield a final concentration of 100 ppm at the output of the olfactometer. During odour delivery, odorants were mixed in air phase by simultaneous opening of all four valves. During the inter-stimulus-interval, air was passed through the blank odour vial containing only DPG. Mice were not subject to this protocol on the day of imaging.

Measuring changes in cortical odour representations after mixture exposure

Cortical representations of all 22 odours of the tiled odour set were obtained after cessation of behavioural training. Single neuron and population representations of the target aldehydes (propanal and butanal) and target ketones (propanone and butanone) were compared to data obtained from odour-naive mice exposed to the tiled odour set. Off-target control comparisons were made between chain-length matched esters and acids (ethyl acetate and propyl acetate versus propanoic acid and butanoic acid) as well as between target ketones and off-target long-chain aldehydes (heptanal, octanal).

For comparing changes in response profiles of individual neurons across the target classes, a class preference index was assigned to each neuron using ROC binary classification. The trial-averaged responses of each neuron were labelled as either aldehydes or ketones, resulting in a single auROC value reflecting the degree of discriminability between classes. The class preference index was obtained by rescaling auROC values from a range of 0 to 1 to a range of −1 to 1 with 0 representing low discriminability. Because the class preference index combines both the magnitude and frequency of responses, a neuron with weak preference for a single class could be strongly and uniformly responsive or non-responsive to the target odours. We therefore limited analysis to neurons exhibiting at least one response to any short-chain aldehyde or ketone of magnitude greater than 2 s.d. above the mean response across all 22 presented odours. Changes to odour similarity at the population level were assessed by comparing the correlation distance between odour pairs across classes. Because differences in the average magnitude of ensemble correlations (which are not relevant to the pairwise restructuring we focus on herein) may uniformly bias comparisons between experimental conditions, before computing pairwise correlation distances (see 'Chemical and activity distance'), the trial-averaged tuning profile of each neuron was initially centred on its mean).

Descriptor relevance

The chemical descriptors in Supplementary Table 1 identified by Lasso optimization afford an algorithmic representation of chemical structure. However, each descriptor incorporates some information about semantic molecular properties, such as molecular weight, electronegativity, polarizability, ionization potential, molecular volume and hydrophobicity. The relevance of each descriptor, obtained from the (Dragon, KODE Inc.) website https://chm.kode-solutions.net/products_dragon_descriptors.php, is presented next to the descriptor name.

Statistical tests

For comparing two normal independent distributions, the Student's *t*-test (two-sided) was used. For comparing two independent distributions when normality cannot be assumed, significance was assessed by permutation testing or using the two-sided Wilcoxon rank sum test. The Kolmogorov–Smirnov test was used to determine equivalence between two distributions. For testing significance of a single statistic against null distributions obtained by permutation, the true value was deemed significant if it resided outside the 5th–95th percentile of the null statistic distribution. Error bars refer to 95th confidence interval, s.e.m. or s.d. as indicated in the figure legends. For regression modelling, confidence intervals are computed over bootstraps (with replacement) of the data. For establishing correspondence between two distance matrices, Pearson's product moment correlation was used on the upper diagonal of each set of measurements.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data will be posted to Github or made available upon reasonable request (www.github.com/dattalab).

Code availability

All code will be posted to Github or made available upon reasonable request (www.github.com/dattalab).

38. Challis, R. C. et al. Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nat. Protoc.* **14**, 379–414 (2019).
39. Bruno, R. M. & Sakmann, B. Cortex is driven by weak but synchronously active thalamocortical synapses. *Science* **312**, 1622–1627 (2006).
40. Minamisawa, G., Funayama, K., Matsuki, N. & Ikegaya, Y. Intact internal dynamics of the neocortex in acutely paralyzed mice. *J. Physiol. Sci.* **61**, 343–348 (2011).
41. Simons, D. J. & Carvell, G. E. Thalamocortical response transformation in the rat vibrissa/barrel system. *J. Neurophysiol.* **61**, 311–330 (1989).
42. Maklad, A., Quinn, T. & Fritzsch, B. Intracranial distribution of the sympathetic system in mice: DiI tracing and immunocytochemical labeling. *Anat. Rec.* **263**, 99–111 (2001).
43. Doevendans, P. A. J., Daemen, M. J., de Muinck, E. D. & Smits, J. F. Cardiovascular phenotyping in mice. *Cardiovasc. Res.* **39**, 34–49 (1998).
44. Carey, R. M. & Wachowiak, M. Effect of sniffing on the temporal structure of mitral/tufted cell output from the olfactory bulb. *J. Neurosci.* **31**, 10615–10626 (2011).
45. Kepecs, A. & Uchida, N. The sniff as a unit of olfactory processing. *Chem. Senses* **31**, 167–179 (2006).
46. Cheung, M. & Carey, R. A method for generating natural and user-defined sniffing patterns in anesthetized or reduced preparations. *Chem. Senses* **34**, 63–76 (2009).
47. Moldestad, O., Karlsen, P., Molden, S. & Storm, J. F. Tracheotomy improves experiment success rate in mice during urethane anesthesia and stereotaxic surgery. *J. Neurosci. Methods* **176**, 57–62 (2009).
48. Ecker, A. S. et al. State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248 (2014).
49. Goard, M. & Dan, Y. Basal forebrain activation enhances cortical coding of natural scenes. *Nat. Neurosci.* **12**, 1444–1449 (2009).
50. Ma, L. et al. Distributed representation of chemical features and tunotopic organization of glomeruli in the mouse olfactory bulb. *Proc. Natl. Acad. Sci. USA* **109**, 5481–5486 (2012).
51. Srinivasan, S. & Stevens, C. A quantitative description of the mouse piriform cortex. Preprint at https://www.bioRxiv.org/content/ 10.1101/099002v1.full (2017).
52. Hagiwara, A., Pal, S. K., Sato, T. F., Wienisch, M. & Murthy, V. N. Optophysiological analysis of associational circuits in the olfactory cortex. *Front. Neural Circuits* **6**, 18 (2012).
53. Luna, V. M. & Morozov, A. Input-specific excitation of olfactory cortex microcircuits. *Front. Neural Circuits* **6**, 69 (2012).
54. Pachitariu, M., Stringer, C., Dipoppa, M. & Schröder, S. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. Preprint at https://www.bioRxiv.org/content/10.1101/061507v2 (2017).
55. Willmore, B. & Tolhurst, D. J. Characterizing the sparseness of neural codes. *Network* **12**, 255–270 (2001).
56. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://arXiv.org/abs/1802.03426 (2018).
57. Haxby, J. V. et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
58. Cowley, B. et al. Distance Covariance Analysis. *Proc. 20th International Conference on Artificial Intelligence and Statistics* **54**, 242–251 (2017).
59. Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. F. Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).
60. Abbott, L. F., Rajan, K. & Sompolinsky, H. Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. Preprint at https://arXiv.org/abs/0912.3832 (2009).
61. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classifiers* **10**, 61–74 (1999).
63. Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
64. Wilson, D. A. & Linster, C. Neurobiology of a simple memory. *J. Neurophysiol.* **100**, 2–7 (2008).

Acknowledgements We thank members of the Datta laboratory, J. Pillow, R. Axel, L. Abbott, A. Litvin-Kumar, C. Schoonover, A. Fink and V. Ruta for comments on the manuscript, S. Knemeyer for technical illustrations, and N. Bhagat for laboratory assistance. We thank N. Mathur for assistance with viral delivery of GCaMP6s. We thank A. Giessel for early development of tools for image acquisition and analysis. We thank O. Mazor and P. Gorelik from the Research Instrumentation Core Facility and Ludo Cacheux for engineering support. Core facility support is provided by NIH P30 grants HD18655 and NS072030. S.R.D. is supported by fellowships from the Vallee Foundation, by grants RO11DC016222 and U19 NS112953 from the National Institutes of Health and by the Simons Collaboration on the Global Brain. S.P. is supported by grant NS108410 from the National Institutes of Health. D.C. is supported by a Bertarelli Foundation Fellowship. S.L.P. is supported by grant DA036922 from the National Institutes of Health and the Stuart H.Q. and Victoria Quan Fellowship. G.I. is supported by the Armenise Foundation Career Development Award.

Author contributions S.L.P. and S.R.D. conceived and designed the experiments. S.L.P. performed imaging experiments and K.D. performed behavioural experiments. G.I. assisted with analysis and performed single-unit recordings to validate the TeLC experiment. D.B. generated reagents and performed infections and histology for the bulbar afferent experiments, and helped to modify and test reagents for the TeLC experiment. K.F. provided reagents and guidance for the TeLC experiment. S.L.P., G.I. and D.C. analysed the data. S.P. provided guidance about population analysis. S.L.P., G.I. and S.R.D. wrote the manuscript.

Competing interests The authors declare no competing interests.

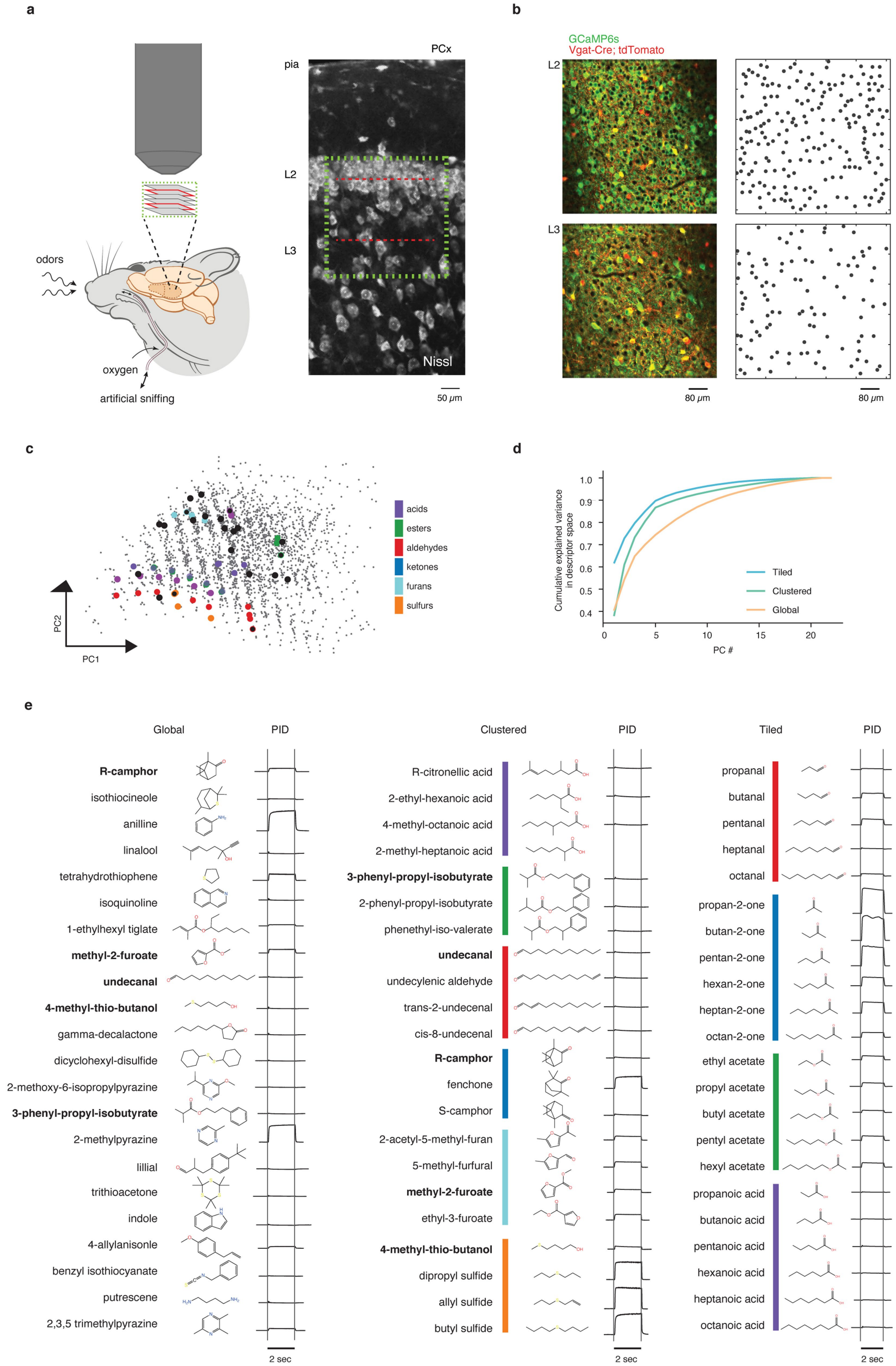
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2451-1.

Correspondence and requests for materials should be addressed to S.R.D.

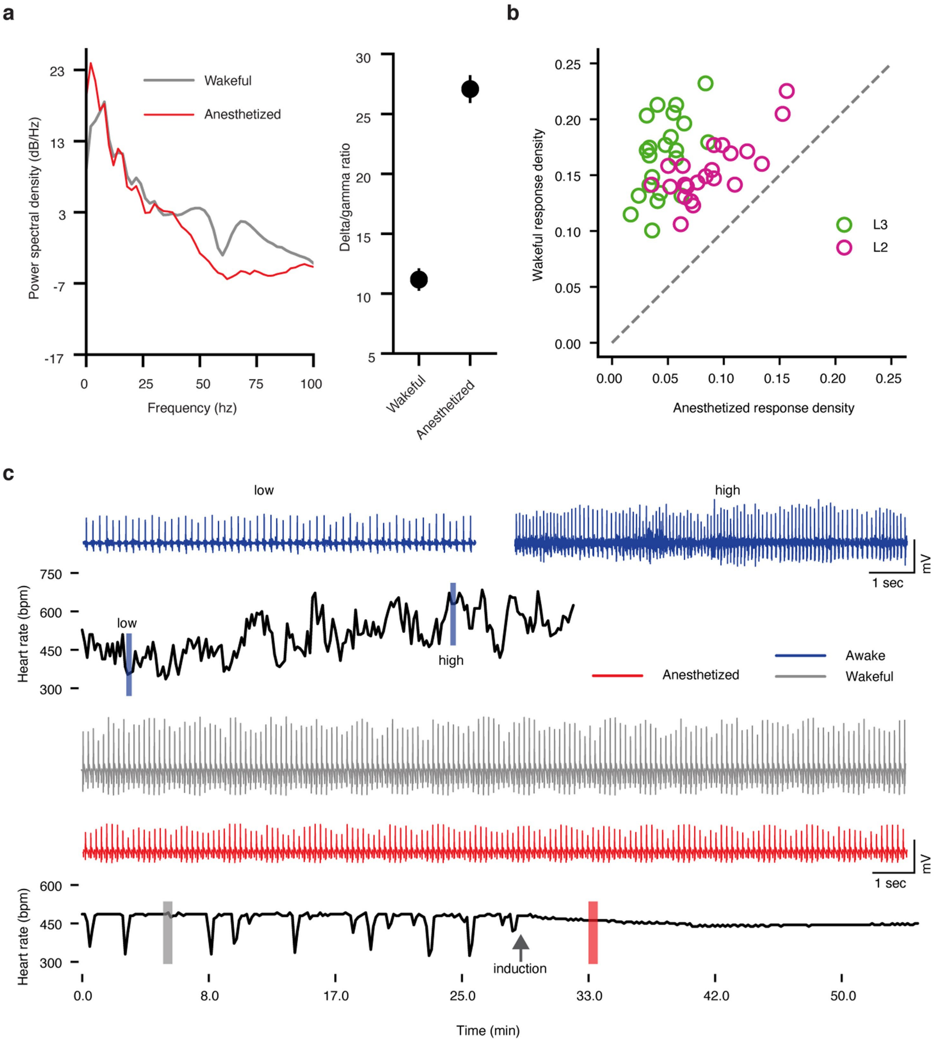
Peer review information *Nature* thanks Vijay Balasubramanian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

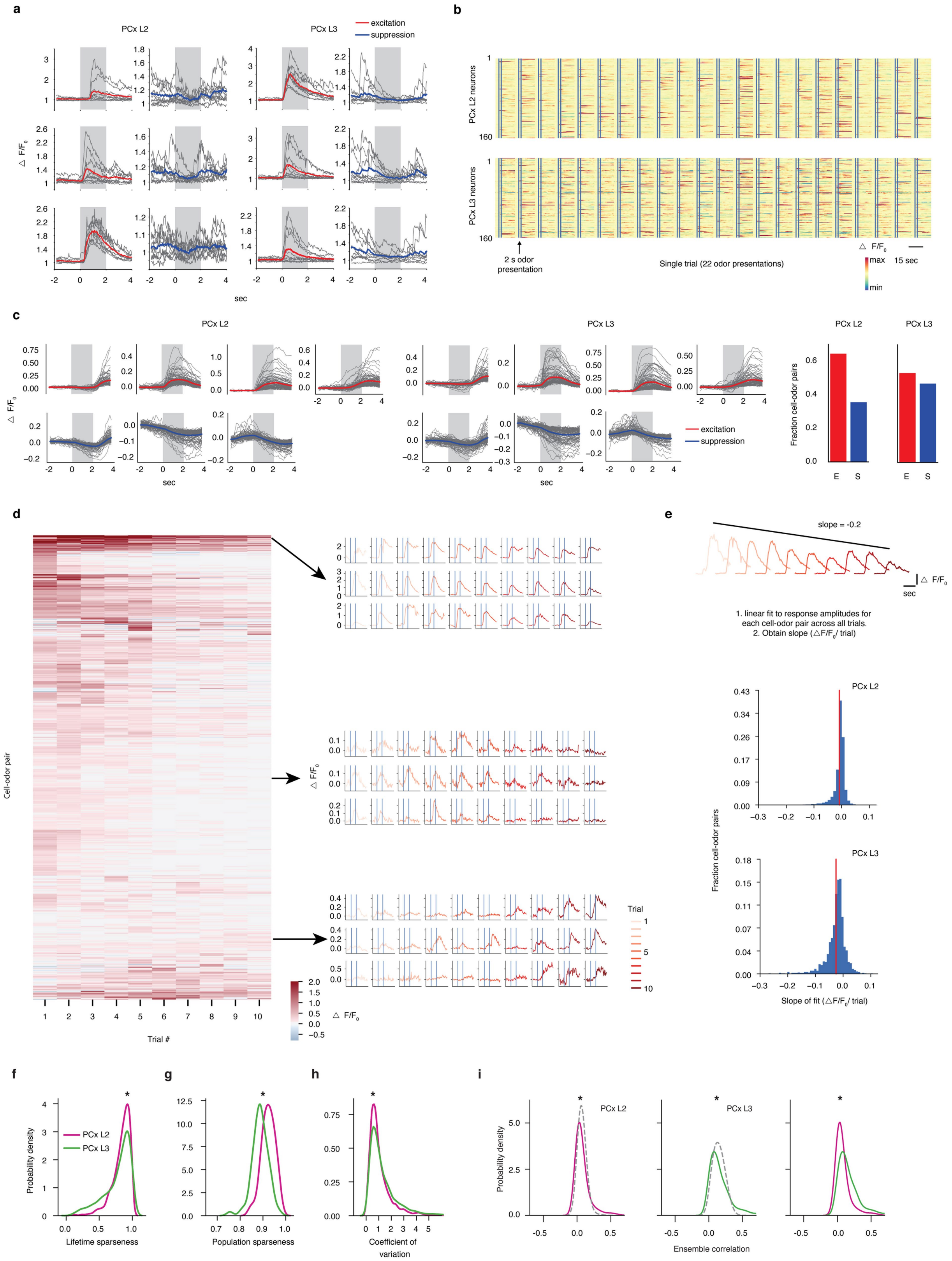


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Volumetric population imaging of PCx L2 and L3 during wakefulness using rationally designed odour sets. a, Left, cartoon of the volumetric multi-photon imaging approach used to characterize odour responses in PCx in wakeful, semi-paralysed mice (Methods). Right, approximate position of an imaging volume (green dotted line) in a typical experiment superimposed on a Nissl-stained coronal section through PCx. Scanning volumes were oriented to acquire similarly sized cortical populations in L2 and L3 (red dotted lines), despite decreased neuron density in L3 (Methods). Imaging was performed in the most anterior portion of the posterior PCx. **b**, Sample fields of view for a single imaging session. PCx L2 is depicted on top; PCx L3 on bottom. Segmentation masks associated with each layer are shown on the right. **c**, Global, clustered and tiled odour sets superimposed on the collection of odours constituting odour space as defined by principal components analysis (Methods). Global odours are indicated by black dots; tiled and clustered odour sets via the indicated colour code. **d**, Plot of the amount of molecular variance contributed by each additional principal component for each odour set in descriptor space; this analysis reveals that each odour set tiles odour space at a distinct level of resolution. **e**, Molecular structures and associated photoionization detector (PID) signals of the odours comprising the global, clustered and tiled odour sets. These PID traces are shown to illustrate the controlled kinetics of the olfactometer only; because detector reports depend on the ability of an odour to be photo-ionized, the relative amplitudes of the traces between odours are not meaningful. For example, heavy aliphatics elicit a minimal PID response because their photo-ionization energies lie outside the range of the detector; however, odours with low or absent PID traces still induced cortical activity in 5–20% of the imaged population, consistent with effective odour delivery. Five odours are shared between the global and clustered odour sets. These are indicated by bold lettering (and in **c**, as black circles with coloured edges). Colour code as in **c**.

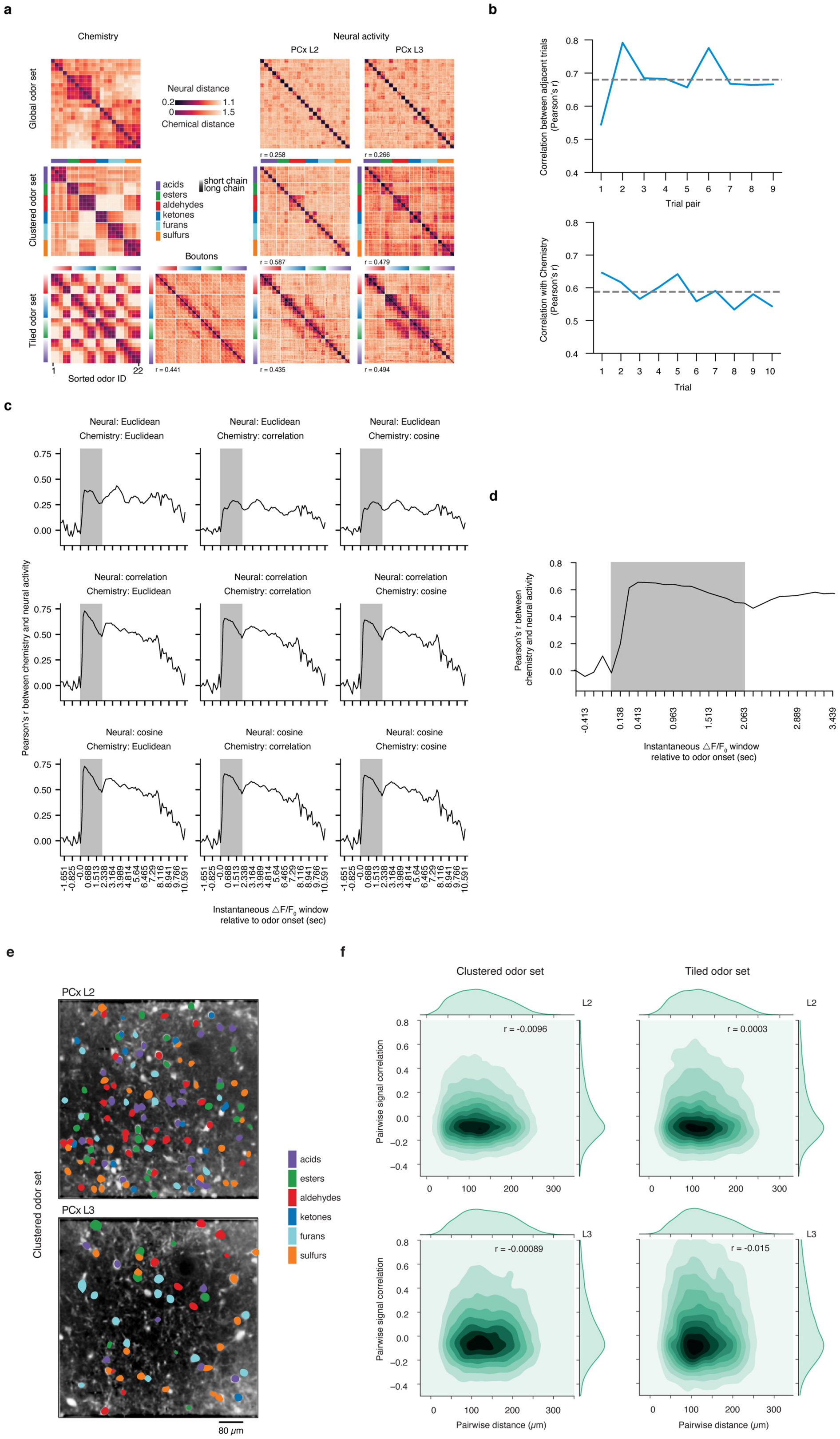


Extended Data Fig. 2 | Odour responses in PCx are substantially altered by anaesthesia. a, Left, EEG power spectral density plot from an individual subject depicting differences in cortical state between ketamine–medetomidine anaesthesia and wakefulness (Methods). Under anaesthesia, the EEG signal is enriched in the delta band (0.5–4 Hz) at the expense of high frequency (40–100 Hz) gamma oscillations; by contrast, gamma activity increases and delta activity decreases during wakefulness. Right, summary of differences in EEG power content expressed as delta/gamma ratio during anaesthesia and wakefulness averaged from four subjects. Error bars indicate s.e.m. **b**, Comparison of the fraction of responsive neurons (obtained from the population of neurons that respond to at least one odour during the wakefulness) (Methods) to the tiled odour set in the same field of view (obtained from PCx L2 and PCx L3) during the awake state and under anaesthesia. Responses are defined according to auROC analysis (Methods). Each dot represents a single odour (L2: 504 neurons, L3: 418 neurons). **c**, Top, black trace represents heart rate (average over 10 s, non-overlapping windows) recorded from an awake mouse in the home cage. Blue traces are example raw heart rate (HR) signal indicating the range of heart rate fluctuations observed during the awake state. The high variability in heart rates (which span approximately 350 to 650 beats per min) reflects ongoing behaviour in the awake mouse. Bottom, as in the top panel, but for heart rate recorded during wakefulness and after induction of ketamine–medetomidine anaesthesia (Methods). Grey arrow indicates time of induction. Grey and red rectangles and associated inset traces are 20-s segments of real-time heart-rate signal. During wakefulness, fluctuations in heart rate remain within a physiologically normal range of 300–500 beats per min, without any detectible episodes of tachycardia (Methods). Periodic large-amplitude dips in the recorded heart rate during wakefulness reflect moments when pharmacological agents are being administered, which briefly interrupts the heart rate monitor.



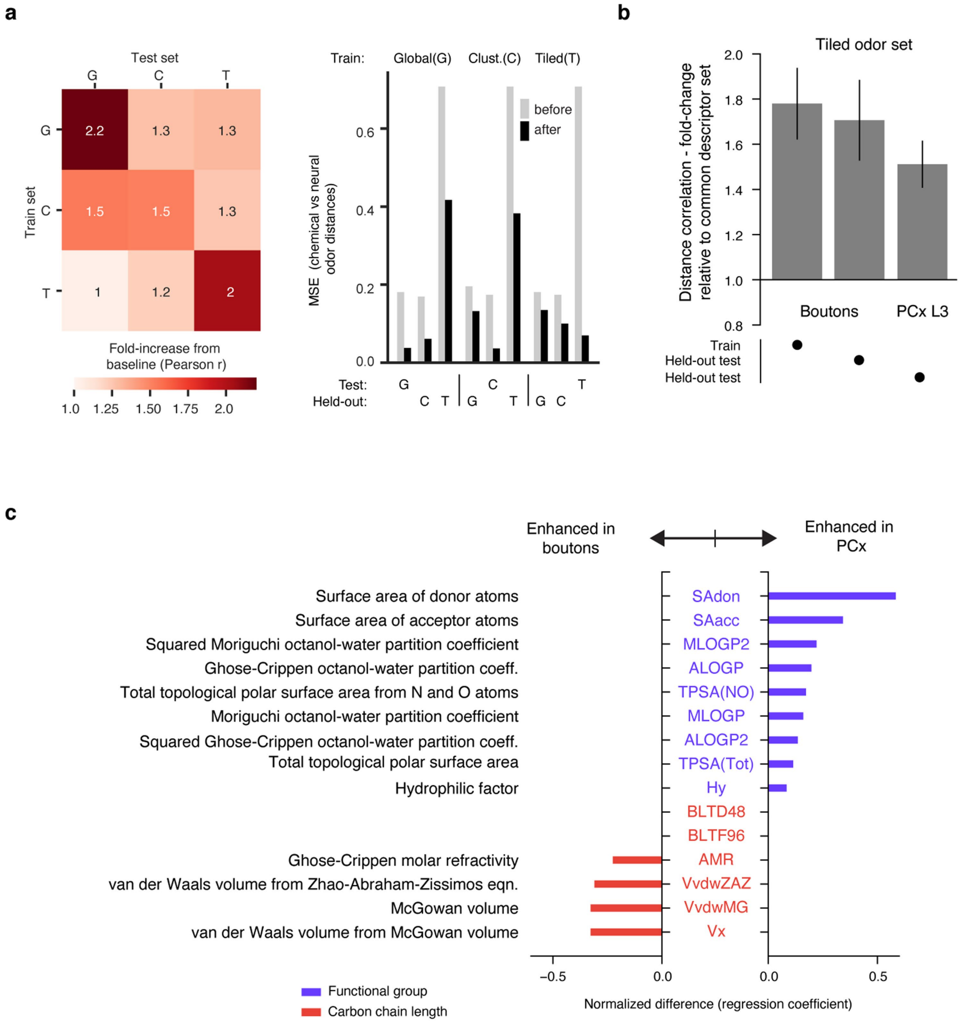
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | PCx L3 neurons exhibit denser, broader and more reliable odour responses than neurons in PCx L2. a, Examples of odour-evoked excitation and suppression in PCx. Each panel corresponds to a single cell-odour pair. Grey lines represent individual trials. Coloured overlays represent trial-mean activity. Shaded grey rectangles delimit the odour presentation period. **b**, Trial-averaged population response raster depicting odour-evoked activity in response to 22 odours (global odour set) across L2 and L3. Responses are $\Delta F/F_0$ with redder colours indicating excitatory transients and bluer colours indicating odour-evoked suppression. *x* axis is time; double vertical bars delimit 2-s odour presentation periods. **c**, Response types observed in L2 and L3 (clustered odour set). Individual panels correspond to clusters identified using a Gaussian mixture model (Methods). Grey traces correspond to trial-averaged cell-odour pairs. Coloured overlays represent mean response time course associated with each cluster. Right, fraction of all cell-odour pairs exhibiting excitation or suppression. **d**, Response amplitudes of cell-odour pairs obtained from PCx L3 depicted on a trial-by-trial basis. Each row represents the response of a given neuron to 10 consecutive presentations of the same odour. Neurons are sorted hierarchically using average linkage and correlation distance. Despite the presence of some habituation in response to several presentations of the same odorant across the experiment, habituation does not appear uniform across the neural population nor does it appear to dominate neural responses to odours. Different groups of neurons were identified with maximal responses to an odour peaking at different times across the experiment; see examples depicted on the right. Each row of traces corresponds to a single cell-odour pair. **e**, At the population level, odour responses do not uniformly habituate across the experiment. Top, cartoon depiction of procedure for determining change in response amplitude over the course of the experiment for a single cell odour pair. Middle and bottom, pooled data for all cell–odour pairs, sorted by layer. Red lines correspond to distribution means (clustered odour set). **f**, Lifetime sparseness distributions (used to quantify tuning breadth) (Methods) in L2 and L3 across all experiments (1 = perfectly odour selective, 0 = completely non-selective, $^*P < 0.01$, permutation test on layer label). Distributions are built using all responsive neurons (significant response to at least one odour by auROC analysis) pooled by layer across all experiments (here and throughout, global: $n = 3$ mice, L2 = 854 neurons, L3 = 616 neurons; clustered: $n = 3$ mice, L2 = 867 neurons, L3 = 488 neurons; tiled: $n = 3$ mice, L2 = 427 neurons, L3 = 334 neurons). **g**, Population sparseness distributions (used to quantify response density) (Methods) in L2 and L3 (1 = few neurons active overall, 0 = all neurons active overall to an equal level). $^*P < 0.01$, permutation test on layer label. **h**, Probability density distributions of coefficient of variation for all significant cell-odour pairs identified with auROC analysis. $^*P < 0.01$, permutation test on layer label. **i**, Probability density distributions of ensemble correlations (that is, pairwise correlations between odour-evoked ensembles) between trial-averaged population odour responses in L2 (left) and L3 (middle). Dashed control curves indicate the distribution of ensemble correlations after shuffling odour labels independently across neurons. Ensemble correlations were determined independently for each mouse, and subsequently pooled. $^*P < 0.01$, permutation test on odour label. L3 exhibits greater correlations at the population level than L2 (right). $^*P < 0.01$, permutation test on layer label.

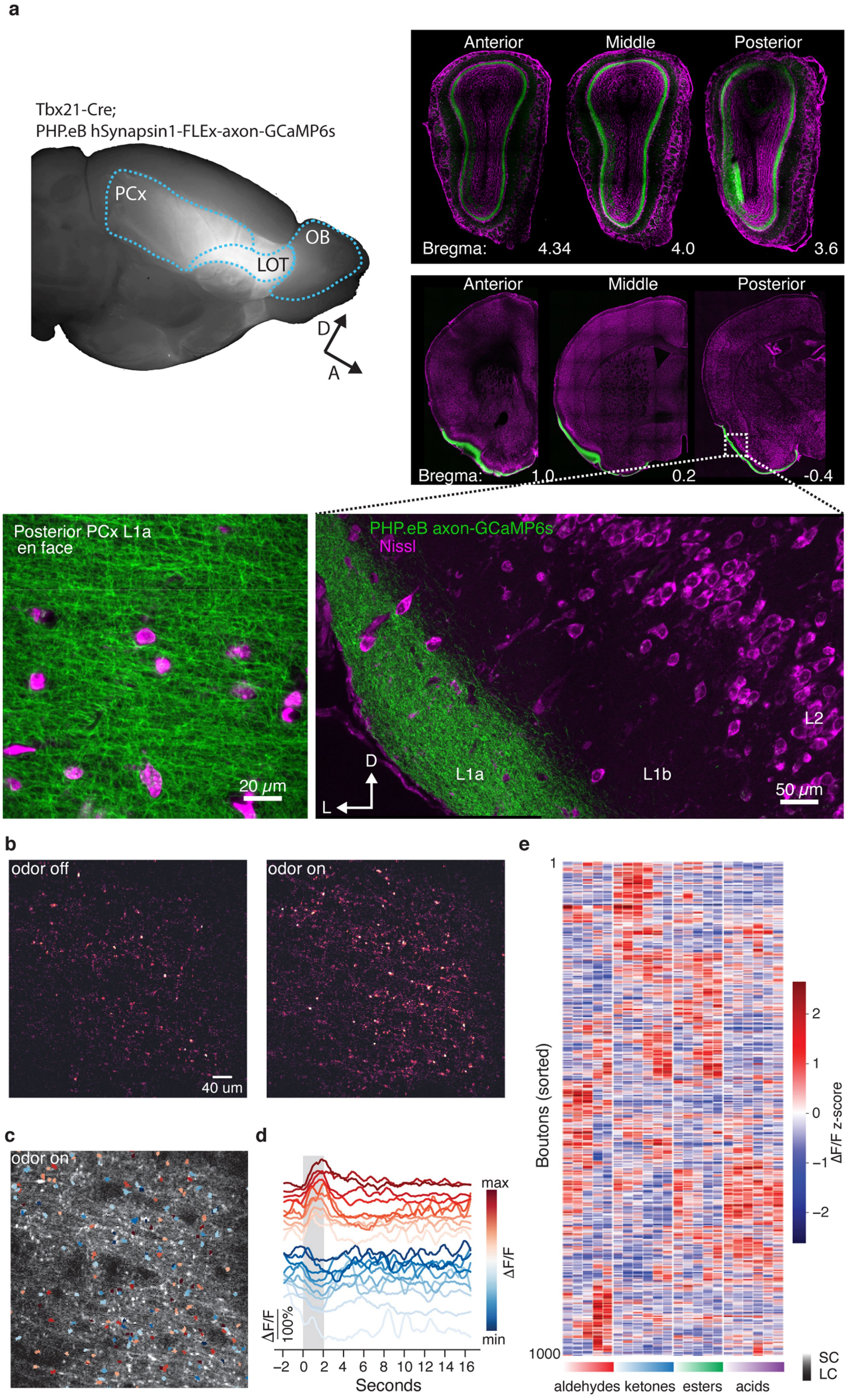


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Cortical odour representations are stable from trial to trial and not chemotopically organized. **a**, Left, pairwise odour chemical correlation matrices for the global, clustered and tiled odour sets. Rows and columns are sorted according to the chemical similarity between odours as assessed by hierarchical clustering (Methods). Middle and right, Pairwise correlation distances of single-trial, population representations for odours in the global, clustered, and tiled odour experiments in PCx L2 and L3 (and boutons for the tiled odour set). Rows and columns are sorted according to the chemical similarity between odours as on the left. Chemical colour code (*x* and *y* axis labels of matrices, indicating functional group associated with each group of molecules) is shown in the legend. *R* values indicate Pearson's correlation to odour chemistry. **b**, Top, structured odour relationships persist from trial to trial over the course of the experiment. Blue line represents the similarity of two correlation distance matrices built from population responses obtained on consecutive trials. Grey dashed line indicates mean across all trial-pair comparisons (10 trials, 9 trial pairs; clustered odour set, L3). Bottom, chemistry-based odour relationships correspond to matched cortical relationships obtained on a trial-by-trial basis. Dashed grey line represents the similarity of chemical and neural activity distances on a trial-by-trial basis. **c**, Correspondence between odour structure in PCx L3 (clustered odour set) and odour chemistry using three different distance metrics (correlation distances, Euclidean distances and cosine distances). Distance matrices calculated from population activity are obtained using instantaneous $\Delta F/F_0$ over 130 ms increments (F_0: baseline fluorescence averaged over a 1-s sliding window). Vertical lines delimit the 2-s odour presentation. **d**, Odour chemical relationships emerge within a few hundred milliseconds after odour onset and persist for several seconds after odour offset (see Extended Data Fig. 1e for associated PID traces). **e**, Example PCx L2 and L3 FOVs from a single mouse with each responsive neuron coloured according to its preferred odour in the clustered odour set. Neurons preferring odours belonging to different classes (legend) appear spatially intermingled in both L2 and L3. **f**, Contour plots of pairwise signal correlations, plotted with respect to distance in L2 and L3 for the clustered and tiled experiments. Darker colours indicate increased density (see margin distributions). Pearson's *r* is overlaid and indicates no spatial organization of odour representations in PCx.

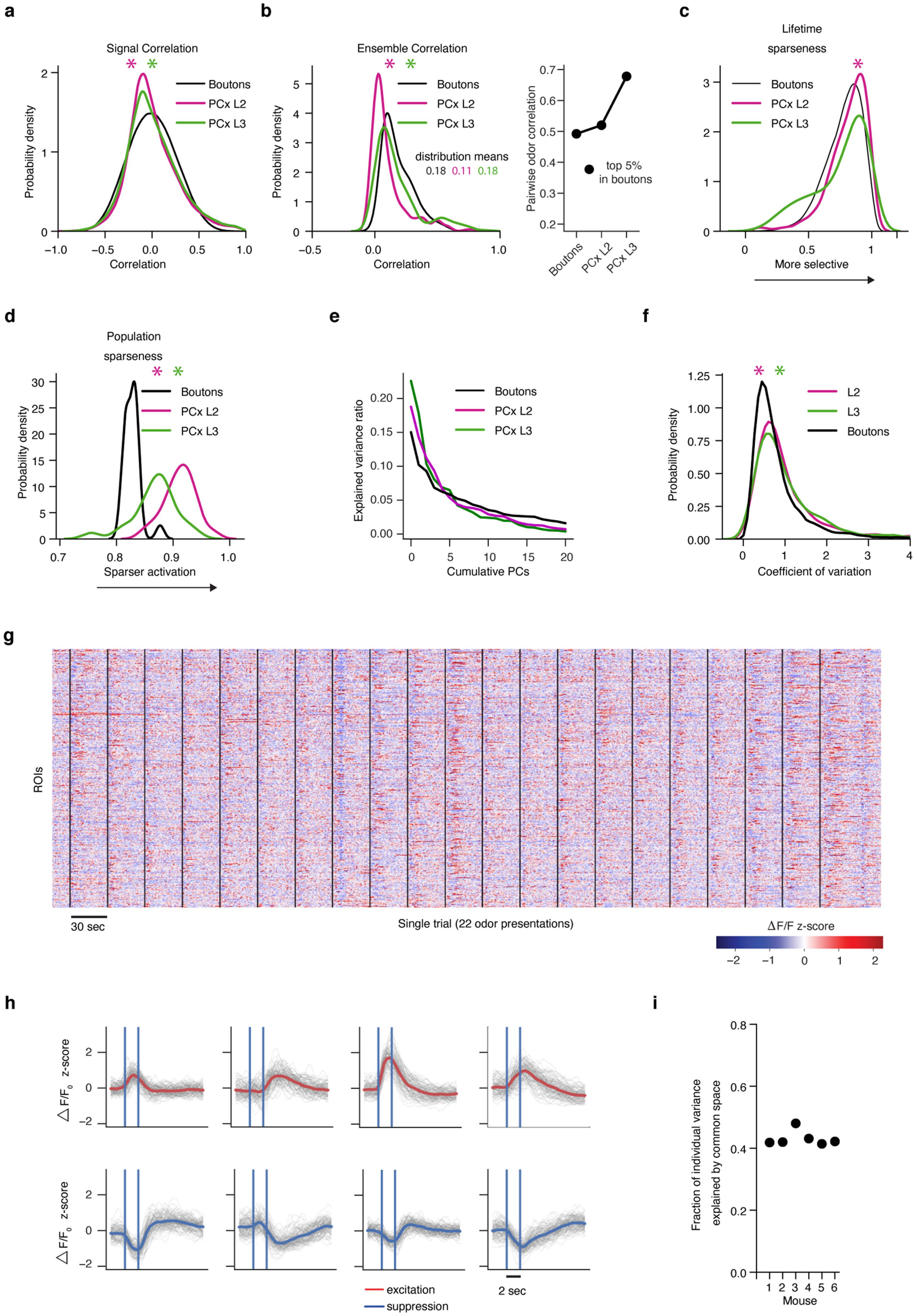


Extended Data Fig. 5 | Lasso optimization identifies parsimonious sets of chemical descriptors that predict neural odour relationships. a, Left, descriptors identified through training on one odour set also improve Pearson's correlation (*r*) between corresponding chemical and neural distances for held-out sets of odours. C, clustered; G, global; T, tiled. A value of 1 in the matrix corresponds to no improvement from baseline Pearson's *r* value after optimization. Baseline chemical-neural correlation is 0.22 for global; 0.48 for clustered; 0.37 for tiled (see Supplementary Table 1 for optimal descriptor sets). Right, reduction in mean-squared error (MSE) between chemical and neural odour pair distances for held-out odour sets (indicated below the *x* axis) after training on a single odour set (indicated above). Note that the five odours in common between the global and clustered odour sets (names in bold in Extended Data Fig. 2e) were discarded when evaluating performance on held-out data. The chemical features learned from the tiled odour set improved chemical–neural Pearson's correlations in the clustered odour experiment but not the global odour experiment, consistent with the odours belonging to the tiled set covering only a limited region of chemical odour space (left). However, despite the limited chemical overlap between the tiled and global odour sets, training on the tiled odour set still improved the correspondence between odour chemistry and neural responses for the global odour set as assessed by a reduction in the mean-squared error (right). **b**, Identifying a subset of chemical descriptors (from the original superset used to define odour space) using Lasso optimization on odour distances improves the correspondence to cortical activity (Methods, Supplementary Table 1). Training data were derived from the bouton dataset, and testing was performed for bouton responses to held-out odours within the tiled odour set, and also to cortical responses of the tiled odour set. Data are mean ± s.e.m. over cross-validation folds. **c**, The same procedure as in **b** was performed on a limited subset of 15 semantically relevant descriptors that comprise the 'molecular properties' block of the Dragon database; these descriptors include metrics that reflect molecular properties associated with functional groups (for example, donor or acceptor atom surface area), molecular weight (for example, van der Waals molecular volume) or a combination of both, such as 'hydrophilic factor', and reflect the main axes of diversity in the tiled odour set. Most descriptors enriched in the olfactory bulb covary with molecular weight (red descriptors). Most descriptors enriched in PCx reflect the combined presence of a charged atom and variable number of carbon atoms along the aliphatic series of the tiled odour set (blue descriptors). Note that these descriptors differ from those identified when querying the entire Dragon set using Lasso optimization (Supplementary Table 1), as this limited set of targeted descriptors (selected because their semantic meaning is transparent) may not afford optimal predictions over neural data.



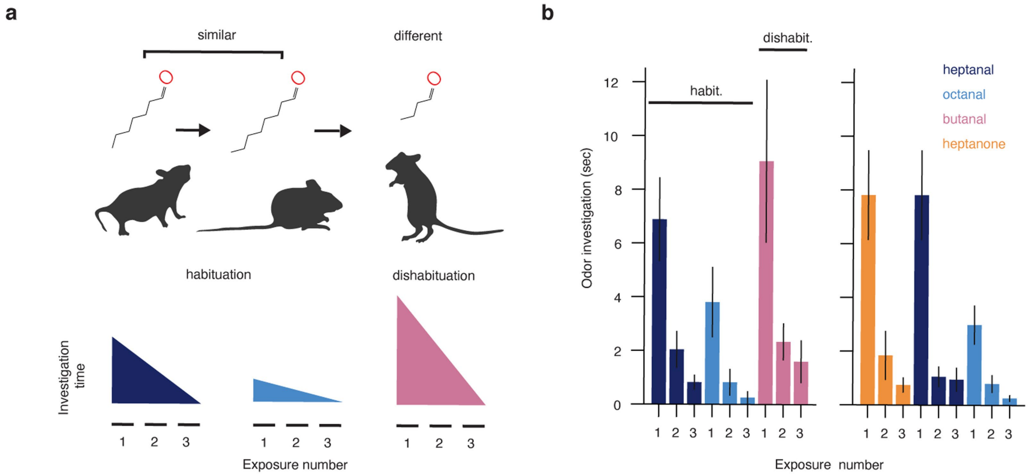
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Functional imaging of OB axons in PCx via axonally targeted GCaMP6s. a, Left, whole-mount depicting Tbx21-Cre-dependent expression of AAV PHP.eB hSynapsin1-FLEX-axon-GCaMP6s in OB projection neuron axons. GCaMP6s fluorescence is broadly distributed across piriform cortex. Right, coronal sections depicting GCaMP6s signal (green) in the mitral cell layer across the entire anterior-posterior extent of the olfactory bulb and cortex. Inset, bottom, GCaMP6s-labelled axons shown coursing through PCx L1a. Bottom left, en face image of L1a depicts dense and uniform distribution of axonal boutons. **b**, Difference heat map of a typical field-of-view (FOV) depicting baseline and odour-driven fluctuations in GCaMP6s signal. The strongest activation (light colour) is associated with axonal boutons. **c**, Time-averaged fluorescence signal of FOV in **b**. Overlay shows segmented ROIs corresponding to axonal boutons depicting increases (red) or decreases (blue) in fluorescence, averaged over multiple presentations of a single odour from the tiled odour set. **d**, Example average fluorescence from several boutons in **a**. Grey bar indicates odour delivery period, scale bar indicates response amplitude. For clarity, fluorescence time courses for each example bouton are offset along the *y* axis. **e**, Example bouton responses for the tiled odour set. Each row represents the trial-averaged response of a single bouton for two seconds during and after odour exposure (columns) depicted as *z*-scored $\Delta F/F_0$; rows are sorted hierarchically using correlation distance and average linkage. The functional group and carbon chain-length associated with each odour are indicated below each column; light-to-saturated gradient indicates progression from short-chain to long-chain odours. Note that, as has been observed previously for OB projection neurons, boutons exhibit a substantial amount of odour-driven suppression.

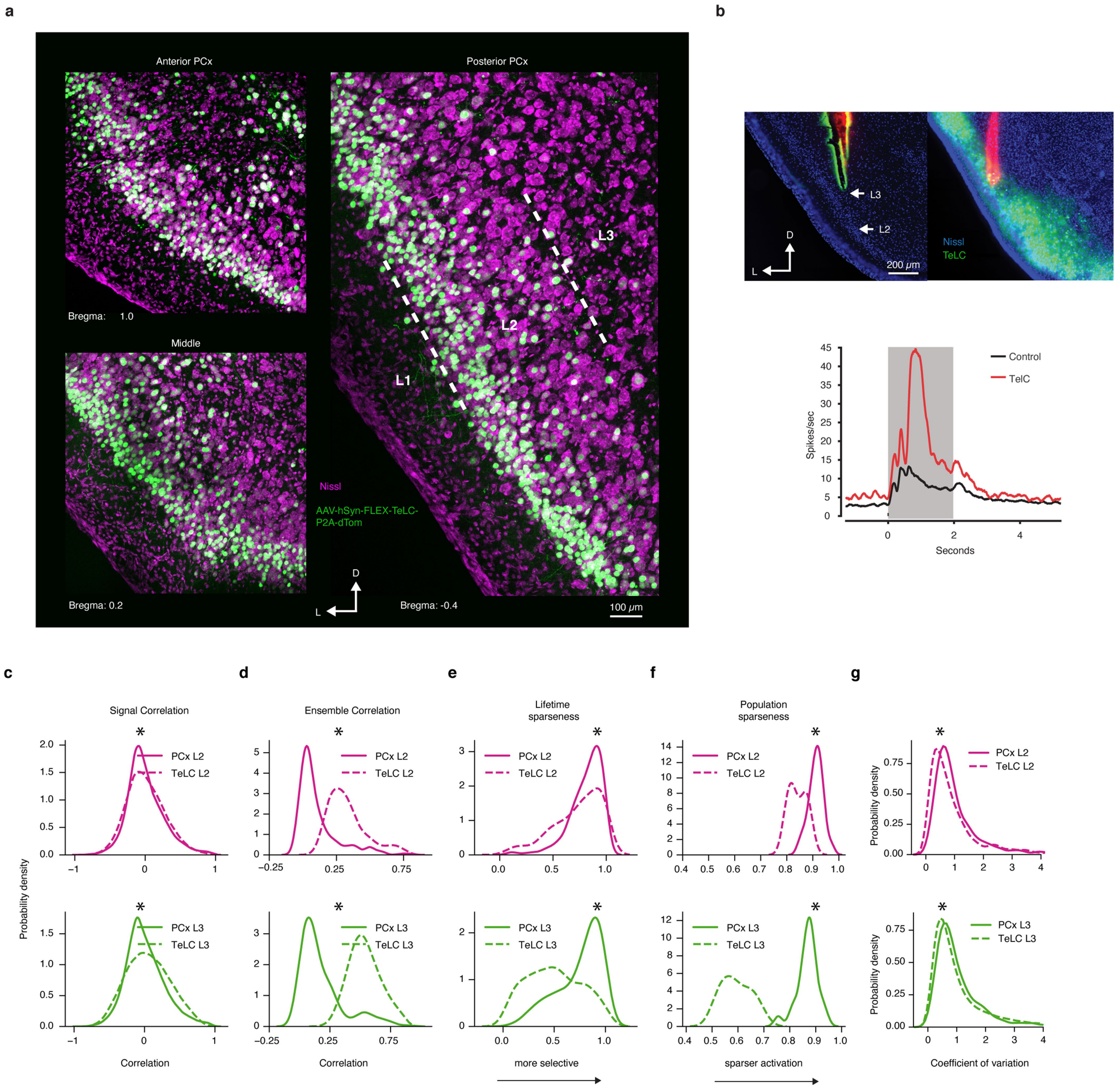


Extended Data Fig. 7 | See next page for caption.

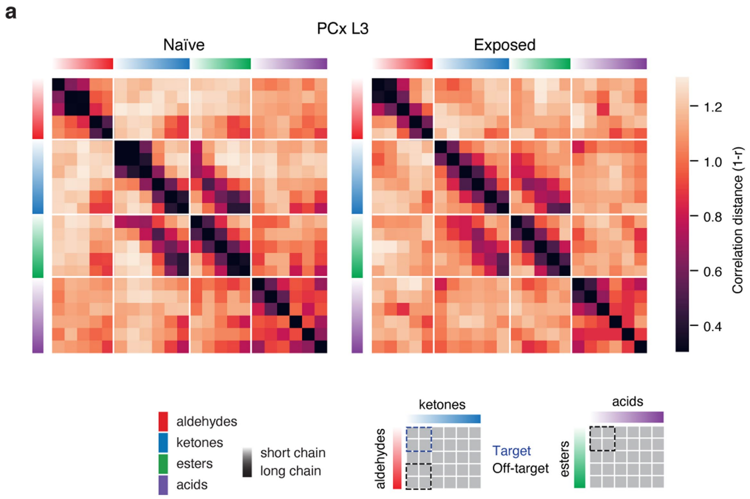
Extended Data Fig. 7 | Bouton odour response properties. **a**, Probability density distributions for boutons, PCx L2, and PCx L3 for signal correlations. **b**, Left, as in **a**, but for ensemble correlations. Right, for the top 5% most similar odour pairs identified in boutons, correlation for the same odour pairs in PCx. Ensemble responses in both PCx L2 and PCx L3 exhibit stronger similarity than boutons. **c**, **d**, Probability density distributions for boutons, PCx L2 and PCx L3, for lifetime and population sparseness. **e**, Cumulative neural variance explained with increasing numbers of principal components, indicating relatively higher dimensionality in boutons compared to PCx (that is, more uniform distribution of variance across principal components). **f**, Probability density distributions for boutons, PCx L2 and PCx L3 for coefficient of variation representing trial-to-trial response variability across cell–odour pairs. These data demonstrate that observed odour responses in boutons are more reliable than similar responses in the cortex. For **a**–**f**, only the tiled odour set is used. For lifetime sparseness, 1 = perfectly odour selective, 0 = completely non-selective. For population sparseness, 1 = few neurons responsive, 0 = all neurons equally responsive. Distributions are built using all responsive neurons/boutons (significant response to at least one odour by auROC analysis; boutons: 3160 ROIs across 6 subjects, PCx L2: 427 neurons across 3 subjects. PCx L3: 334 neurons across 3 subjects). Asterisk indicates significant difference between boutons and either L2 or L3: **a**, vs L2 $P < 10^{-27}$; vs L3 $P = 0.02$; **b**, vs L2 $P < 10^{-20}$; vs L3 $P < 0.005$; **c**, vs L2 $P < 10^{-9}$; vs L3 $P = 0.93$; **d**, vs L2 $P < 10^{-7}$ vs L3 $P < 10^{-4}$; **f**, vs L2: $P < 10^{-20}$; vs L3: $P < 10^{-23}$; two-sided Wilcoxon rank sum test for all comparisons. **g**, Single-trial Z-scored $\Delta F/F_0$ for 1,000 boutons recorded in PCx L1a during presentation of 22 odours belonging to the tiled odour set indicated by black lines. Redder colours indicating excitatory transients, and bluer colours indicate odour-evoked suppression. **h**, Response types observed in boutons (tiled odour set). Individual panels correspond to clusters identified using a Gaussian mixture model (Methods). Grey traces correspond to trial-averaged bouton-odour pairs. Coloured overlays represent mean response time course associated with each cluster. Blue vertical lines mark periods of odour presentation. **i**, Fraction of total odour-driven bouton variance in each individual mouse that can be attributed to the shared across-mouse structure as quantified by distance covariance analysis (Methods).



Extended Data Fig. 8 | Habituation-dishabituation test for assessing perceptual similarity of odour pairs. a, Left, mice presented with new odours exhibit investigation that diminishes over several consecutive presentations of the same odorant. Subsequent presentation of a perceptually different odour reinstates investigation, and presentation of a similar odour has little effect. The extent to which two odorants are perceptually related is assessed by the magnitude of rekindled interest in the second odour after habituation has occurred to the first. **b**, Investigation times for two different odour triplets. Data are mean ± s.e.m. ($n = 7$ and $n = 8$ mice, respectively). After habituation to heptanal, investigation of the closely related octanal (1-carbon difference) only modestly increases. Presentation of butanal after habituation to octanal (4-carbon difference) induces greater investigation. For the second triplet, presentation of heptanal after habituation to heptanone (0-carbon difference, different functional group) induces greater investigation, whereas subsequent presentation of octanal after habituation to heptanal (1-carbon difference, same functional group) induces much less investigation.



Extended Data Fig. 9 | Inhibition of the associative network through cell-autonomous expression of tetanus toxin light chain in excitatory PCx neurons. a, Uniform infection of excitatory pyramidal neurons in PCx L2 and L3 with AAV-hSyn-FLEX-TeLC-P2A-NLS-dTom in an Emx1-Cre mouse. **b**, Left, coronal section through PCx indicating placement of recording electrode. Right, single-unit odour-evoked activity (grand-average of all excitatory responses deemed as significant by auROC analysis) in Emx1-Cre mice expressing TeLC or wild-type controls. Disruption of cortical recurrent excitation enhances odour-evoked excitation, consistent with disruption of feedback inhibition. Grey bar indicates odour presentation ($n = 121$ cell–odour pairs from two Emx1-Cre mice expressing TeLC; $n = 229$ cell–odour pairs from four mice). **c**–**g**, Probability density distributions for the TeLC experiment for signal and ensemble correlations, lifetime and population sparseness, and coefficient of variation (constructed as in Extended Data Fig. 7, here only for the tiled odour set). For lifetime sparseness, 1 = perfectly odour selective, 0 = completely non-selective. For population sparseness, 1 = few neurons responsive, 0 = all neurons equally responsive. Distributions are built using all responsive neurons (significant response to at least one odour by auROC analysis; TeLC L2: 435 neurons across 3 subjects. TeLC L3: 590 neurons across 3 subjects. PCx L2: 427 neurons across 3 subjects. PCx L3: 334 neurons across 3 subjects). Asterisk indicates TeLC is significantly different from PCx L2 or L3: **c**, L2 $P < 10^{-8}$; L3 $P < 10^{-198}$; **d**, L2 $P < 10^{-46}$; L3 $P < 10^{-55}$; **e**, L2 $P < 10^{-05}$; L3 $P < 10^{-37}$; **f**, L2 $P < 10^{-7}$ L3 $P < 10^{-8}$; **g**, L2: $P < 10^{-10}$; L3: $P < 10^{-4}$; two-sided Wilcoxon rank sum test for all comparisons.



Extended Data Fig. 10 | Passive odour experience modifies odour relationships. a, Correlation distance matrices for the tiled odour set obtained from odour-naive (same data as in Figs. 1–4) mice as well as mice passively exposed to a target mixture of two short-chain aldehydes and two short-chain ketones in the home cage (Methods, Fig. 4e, f). Passive experience with the mixture increases odour similarity specifically between mixture components (target comparisons indicated in the legend in blue), but not between target ketones and long-chain aldehydes or short-chain esters and short-chain acids with which mice had no previous experience (off-target comparisons indicated in legend in black, naïve: 334 neurons, $n = 3$ mice; exposed: 742 neurons, $n = 3$ mice).

nature research

Corresponding author(s): Sandeep R. Datta

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Imaging data was acquired using Scanimage 5 by Vidrio. Extraction of cell fluorescence was performed using the open-source software Suite2p. The odor library used for designing stimulus sets was obtained from www.thegoodscentscompany.com. Physicochemical descriptors were calculated using Dragon 7.0, KODE Inc.
Data analysis	Standard analyses were performed in Matlab and Python. The package "Pyrcca" (https://github.com/gallantlab/pyrcca) was used for multi-set canonical correlation analysis. Simulated annealing was performed using the open-source package simanneal 0.4.2 (https://pypi.org/project/simanneal/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Data collected from 18 animals was included in the study: each of three odor sets described in the study was presented to three animals such that each animal was exposed once to a single odor set for piriform imaging, the tiled odor set was presented to 3 additional mice for cortical imaging in the TeLC experiment, and six additional mice for the bouton imaging experiment. We found that due to the consistency in the structure of odor representations apparent across individuals, three animals sufficed for each condition.
Data exclusions	Data exclusion criteria are described in the Methods. Briefly, exclusion criteria were pre-established to select for experiments where imaging volumes spanning both Piriform cortical layers 2 and 3 (or in the case of the bouton experiment, Layer 1a) could be imaged continuously for at least 2.5 hours with minimal drift and motion artifacts, where a sufficient fraction of cortical neurons was labeled with GcAMP6s and where odor-evoked activity could be detected over the course of the entire imaging session.
Replication	The principal findings in this study were present in each individual and are reported in the main text.
Randomization	Not relevant. Animals were not assigned to different conditions based on whether they originated from the same breeding, litter, or housing cage, as all mice were derived from different cages and different breedings. In addition, the consistency in the structure of odor representations described in this study suggests that the observations are robust to these variables.
Blinding	Not relevant. The main findings in this study are not contingent on blind comparison of two or more experimental conditions, except for the scoring of the behavioral experiment, which was carried out in a blinded fashion.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Male 8-16 week old C57/BL6J mice; for cortical imaging, mice harboring the Vgat-ires-Cre (Jackson Stock No. 028862) and ROSA26-LSL-TdTomato reporter alleles (Jackson Stock No. 007914) were used, and for bulb imaging, mice harboring the Tbet-Cre

allele (Jackson Stock 024507) were used.

Wild animals	This study did not involve the use of wild animals.
Field-collected samples	This study did not involve samples collected from the field.

Insights into variation in meiosis from 31,228 human sperm genomes

https://doi.org/10.1038/s41586-020-2347-0

Received: 1 May 2019

Accepted: 23 March 2020

Avery Davis Bell[1,2]✉, **Curtis J. Mello**[1,2], **James Nemesh**[1,2], **Sara A. Brumbaugh**[1,2], **Alec Wysoker**[1,2] & **Steven A. McCarroll**[1,2]✉

Meiosis, although essential for reproduction, is also variable and error-prone: rates of chromosome crossover vary among gametes, between the sexes, and among humans of the same sex, and chromosome missegregation leads to abnormal chromosome numbers (aneuploidy)[1–8]. To study diverse meiotic outcomes and how they covary across chromosomes, gametes and humans, we developed Sperm-seq, a way of simultaneously analysing the genomes of thousands of individual sperm. Here we analyse the genomes of 31,228 human gametes from 20 sperm donors, identifying 813,122 crossovers and 787 aneuploid chromosomes. Sperm donors had aneuploidy rates ranging from 0.01 to 0.05 aneuploidies per gamete; crossovers partially protected chromosomes from nondisjunction at the meiosis I cell division. Some chromosomes and donors underwent more-frequent nondisjunction during meiosis I, and others showed more meiosis II segregation failures. Sperm genomes also manifested many genomic anomalies that could not be explained by simple nondisjunction. Diverse recombination phenotypes—from crossover rates to crossover location and separation, a measure of crossover interference—covaried strongly across individuals and cells. Our results can be incorporated with earlier observations into a unified model in which a core mechanism, the variable physical compaction of meiotic chromosomes, generates interindividual and cell-to-cell variation in diverse meiotic phenotypes.

One way to learn about human meiosis has been to study how genomes are inherited across generations. Genotype data are available for millions of people and thousands of families; crossover locations are estimated from genomic segment sharing among relatives and from linkage-disequilibrium patterns in populations[2,4,7,9,10]. Although inheritance studies sample only the few gametes per individual that generate offspring, such analyses have revealed that average crossover numbers and crossover locations associate with common variants at many genomic loci[3–6,11,12].

Another powerful approach to studying meiosis is to directly visualize meiotic processes in gametocytes, which has made it possible to see that homologous chromosomes usually begin synapsis (their physical connection) near their telomeres[13–15]; to observe double-strand breaks, a subset of which progress to crossovers, by monitoring proteins that bind to such breaks[16,17]; and to detect adverse meiotic outcomes, such as chromosome missegregation[18,19]. Studies based on such methods have revealed much cell-to-cell variation in features such as the physical compaction of meiotic chromosomes[20,21].

More recently, human meiotic phenotypes have been studied by genotyping or sequencing up to 100 gametes from one person, demonstrating that crossovers and aneuploidy can be ascertained from direct analysis of gamete genomes[22–26]. Despite these advances, it has not yet been possible to measure multiple meiotic phenotypes genome-wide in many individual gametes from many people.

Development of Sperm-seq

We developed a method ('Sperm-seq') with which to sequence thousands of sperm genomes quickly and simultaneously (Fig. 1). A key challenge in developing Sperm-seq was to deliver thousands of molecularly accessible-but-intact sperm genomes to individual nanolitre-scale droplets in solution. Tightly compacted[27] sperm genomes are difficult to access enzymatically without loss of their DNA into solution; we accomplished this by decondensing sperm nuclei using reagents that mimic the molecules with which the egg gently unpacks the sperm pronucleus (Extended Data Fig. 1a–d). We then encapsulated these sperm DNA 'florets' into droplets together with beads that delivered unique DNA barcodes for incorporation into the genomic DNA of each sperm; we modified three technologies to do this (Drop-seq[28], 10× Chromium Single Cell DNA, and 10× GemCode[29], the latter of which was used to generate the data in this study) (Extended Data Fig. 1e, f). We then developed, adapted and integrated computational methods for determining the chromosomal phase of the sequence variants of each donor and for inferring the ploidy and crossovers of each chromosome in each cell.

We used this combination of molecular and computational approaches to analyse 31,228 sperm cells from 20 sperm donors (974–2,274 gametes per donor), sequencing a median of roughly 1% of the haploid genome of each cell (Extended Data Table 1). Deeper sequencing allows detection of roughly 10% of a gamete's genome.

---

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
✉e-mail: averydavisbell@gmail.com; mccarroll@genetics.med.harvard.edu

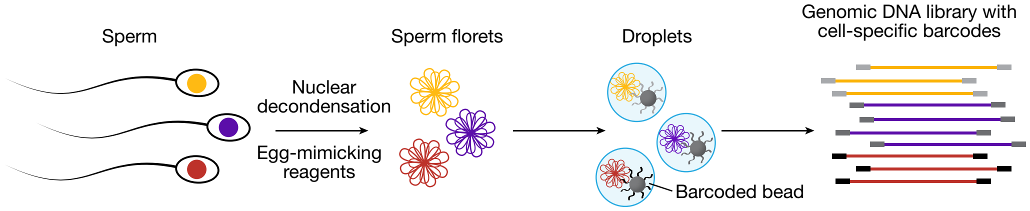


Fig. 1 | Overview of Sperm-seq. Schematic showing our droplet-based single-sperm sequencing method.

Sperm-seq enabled us to infer the haplotypes of donors along the full length of every chromosome: alleles from the same parental chromosome tend to appear in the same gametes, so the coappearance patterns of alleles across many sperm enabled us to assemble alleles into chromosome-length haplotypes (Extended Data Fig. 2a and Methods). In silico simulations and comparisons with kilobase-scale haplotypes from population-based analyses indicated that Sperm-seq assigned alleles to haplotypes with 97.5–100% accuracy (Extended Data Fig. 2b, c and Supplementary Notes).

The phased haplotypes determined by Sperm-seq allowed us to identify cell 'doublets' from the presence of both parental haplotypes at loci on multiple chromosomes (Extended Data Fig. 2d–f and Methods). We also identified surprising 'bead doublets', in which two beads' barcodes reported identical haplotypes genome-wide through different single-nucleotide polymorphisms (SNPs), and thus appeared to have been incorporated into the same gamete genome (Extended Data Fig. 3a, b, Methods and Supplementary Methods). Bead doublets were useful for evaluating the replicability of Sperm-seq data and analyses (Extended Data Fig. 3c–e), which is usually impossible to do in inherently destructive single-cell sequencing.

Recombination rate in sperm donors and cells

We identified crossover (recombination) events in each cell as transitions between the parental haplotypes we had inferred analytically (Methods). We identified 813,122 crossovers in the 31,228 gamete genomes (Extended Data Table 1). Crossover locations were inferred with a median resolution of 240 kilobases (kb), with 9,746 (1.2%) inferred within 10 kb (Extended Data Table 1 and Supplementary Notes). Analysis of bead doublets indicated high accuracy of crossover inferences (Extended Data Fig. 3e). Estimates of crossover rate and location were robust to downsampling to the same coverage in each cell (Extended Data Fig. 4 and Supplementary Methods).

The recombination rates of the 20 sperm donors ranged from 22.2 to 28.1 crossovers per cell. This is consistent with estimates from other methods[3,5,6,10–12,24,26], but with far more precision at the individual-donor level (95% confidence intervals of 22.0–22.4 to 27.9–28.4 crossovers per cell) owing to the large number of gametes analysed per donor (Extended Data Table 1 and Extended Data Fig. 5a). Individuals with higher global crossover rates had more crossovers on average on each chromosome (Extended Data Fig. 5b). We generated genetic maps for each of the donors from their 25,839–62,110 observed crossovers; these maps were broadly concordant with a family-derived paternal genetic map[6] (Extended Data Fig. 5c, d, Supplementary Notes and Supplementary Methods).

Much more variation was present at the single-cell level: cells routinely contained 17 to 37 crossovers (1st and 99th percentiles, median across donors), with a standard deviation of 4.23 across cells (median across donors), versus a standard deviation of 1.53 across donors' crossover rates. Among gametes from the same donor, gametes with fewer crossovers in half of their genome tended to have fewer crossovers in the other half of their genome (Pearson's $r = 0.09$, two-sided $P = 8 \times 10^{-54}$ with all gametes from all donors combined after within-donor normalization) (Supplementary Notes). This relationship, predicted by earlier observations in families[5] and spermatocytes[21], suggests that the crossover number on each chromosome is partly shaped by factors that act nucleus-wide.

Crossover location and interference

All 20 donors shared a tendency to concentrate their crossovers in the same regions of the genome, with large concentrations of crossovers in distal regions, as expected from earlier analyses of families[4,6,9,11,30], and more modest shared enrichments in many centromere-proximal regions (Fig. 2a and Extended Data Fig. 6). Guided by these empirical patterns, we divided the genome into 'crossover zones', each bounded by local minima in crossover density (Extended Data Fig. 6b and Supplementary Methods). These zones are much larger-scale than fine-scale-sequence-driven crossover hotspots[7,31–33], which the spatial resolution of most crossover inferences was not well-suited for analysing.

The crossover zones with the most variable usage across people were all adjacent to centromeres; individuals with high recombination rates used these zones much more frequently (Fig. 2a and Extended Data Fig. 6a; with simulated equal SNP coverage, Extended Data Fig. 4c, e). The relative usage of distal and proximal zones varied greatly among donors and correlated with donors' recombination rates (Extended Data Fig. 7). These results were robust to alternative definitions of 'distal' versus 'proximal' (Extended Data Fig. 7c and Supplementary Notes).

Positive crossover interference causes crossovers in the same meiosis to be further apart than they would be if crossovers were independent events[26,30,34,35]. The effect of crossover interference was visible in each of the 20 sperm donors (Extended Data Fig. 8 and Supplementary Methods). Crossover separation varied greatly among sperm donors and correlated inversely with recombination rate (Extended Data Fig. 7b)—results that were robust to chromosome composition and that applied similarly to same-arm and opposite-arm crossover pairs (Extended Data Fig. 7e, f and Supplementary Notes).

The extremely strong correlations of donors' crossover rates with crossover locations and interference could arise from an underlying biological factor that coordinates these phenotypes, or could arise trivially from the fact that chromosomes with more crossovers would also tend to have crossovers more closely spaced and in more regions. To distinguish between these possibilities, we focused on data from the 180,738 chromosomes with exactly two crossovers (here called 'two-crossover chromosomes') (Supplementary Notes). Even in this two-crossover chromosome analysis, distal-zone usage (Fig. 2b) and crossover separation (Fig. 2c) correlated strongly and negatively with genome-wide recombination rate (additional control analyses are described in the Supplementary Notes and Extended Data Fig. 7d, g, h). These relationships indicate that a donor's crossover-location and crossover-spacing phenotypes reflect underlying biological factors that vary from person to person, as opposed to resulting indirectly from the number of crossovers on a chromosome.

To test whether this covariation of diverse meiotic phenotypes also governs variation at the single-gamete level, we investigated whether cells with more crossovers than the average for their donor also exhibit the same kinds of crossover-spacing and crossover-location

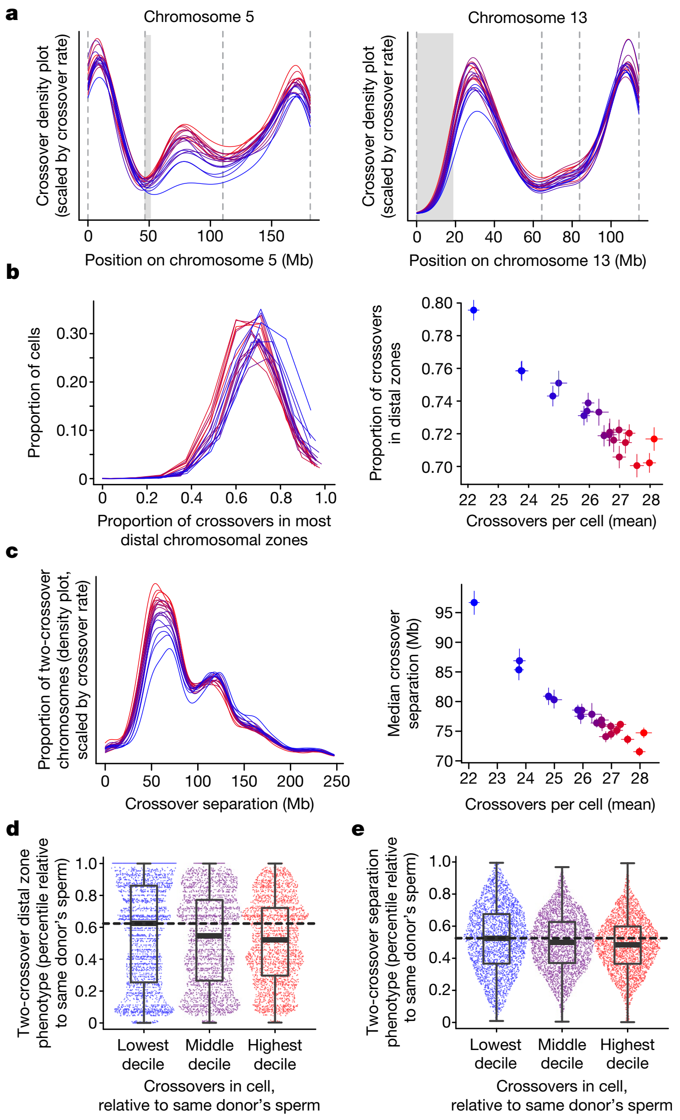


Fig. 2 | Variation in crossover positioning and crossover separation (interference). Colours indicate the crossover rate of donor or cell (blue, low; red, high). **a**, Crossover location density plots for two chromosomes (5 and 13) from each donor ($n = 20$). Dashed grey vertical lines show boundaries between crossover zones. Mb, megabases. **b**–**e**, Crossover positioning and separation (interference) on chromosomes with two crossovers. **b**, **c**, Interindividual variation among $n = 20$ sperm donors. Error bars show 95% confidence intervals. **b**, Left, per-cell proportion of crossovers in the most distal crossover zones (Kruskal–Wallis chi-squared = 1,034; df = 19; $P = 2 \times 10^{-207}$). Right, mean crossover rate versus the proportion of all crossovers (on two-crossover chromosomes) occurring in distal zones (Pearson's $r = -0.95$; two-sided $P = 8 \times 10^{-11}$). **c**, Left, density plot of separation between consecutive crossovers (Kruskal–Wallis chi-squared = 1,792; df = 19; $P < 10^{-300}$). Right, mean crossover rate versus median crossover separation on two-crossover chromosomes (Pearson's $r = -0.95$; two-sided $P = 7 \times 10^{-11}$). **d**, **e**, Among-cell covariation of crossover rate with distal zone use (**d**) or crossover interference (**e**). Phenotypes are analysed as percentiles relative to sperm from the same donor. Box plots: midpoints, medians; boxes, 25th and 75th percentiles; whiskers, minima and maxima. **d**, Single-cell distal-zone use (the proportion of crossovers on two-crossover chromosomes that are in the most distal zones) versus crossover rate (n cells per decile = 3,152, 3,080 and 3,101 for first, fifth and tenth deciles, respectively; Mann–Whitney $W = 5{,}271{,}934.5$; two-sided $P = 2 \times 10^{-9}$ between first and tenth deciles). **e**, Single-cell crossover separation (the median of all fractions of a chromosome separating consecutive two-crossover chromosome crossovers in each cell) versus crossover rate (Mann–Whitney $W = 148{,}548{,}161$, two-sided $P = 3 \times 10^{-53}$ between first ($n = 11{,}658$) and tenth ($n = 23{,}154$) deciles; all intercrossover separations used in test).

phenotypes that donors with high crossover rates do (Supplementary Methods). Indeed, two-crossover chromosomes from cells with more crossovers tended to have closer crossover spacing and increased relative use of non-distal zones (Fig. 2d, e and Extended Data Fig. 7i, j; unnormalized results are in the Supplementary Notes). This result indicates that the correlated meiotic-outcome biases that distinguish people from one another also distinguish the gametes within each individual (see Discussion).

Chromosome and sperm donor aneuploidy

Aneuploidy generally arises from a chromosome missegregation that yields two aneuploid cells: one in which that chromosome is absent (a loss), and one in which it is present in two copies (a gain). Among the 31,228 gametes, we found 787 whole-chromosome aneuploidies and 133 chromosome arm-scale gains and losses (2.5% and 0.4% of cells, respectively) (Fig. 3a and Methods). All chromosomes and sperm donors were affected. The sex chromosomes and acrocentric chromosomes had the highest rates of aneuploidy, consistent with estimates based on fluorescence in situ hybridization analysis of chromosomes[18,19] (Fig. 3b).

The 20 young (18–38-year-old) sperm donors, considered by clinical criteria to have normal-range sperm parameters, exhibited aneuploidy frequencies ranging from 0.010 to 0.046 aneuploidy events per cell (Fig. 3c and Extended Data Table 1). Permutation tests indicated that this 4.5-fold variation in observed aneuploidy rates reflected genuine inter-individual variation (one-sided $P < 0.0001$) (Supplementary Notes).

Under the prevailing model for the origins of aneuploidy, sperm with chromosome losses and gains should be equally common. However, we observed 2.4-fold more chromosome losses than chromosome gains (554 losses versus 233 gains; proportion test two-sided $P = 2 \times 10^{-30}$). This asymmetry did not appear to reflect technical ascertainment bias (Extended Data Fig. 9a and Supplementary Notes). This result is considered further in the Supplementary Discussion.

Errors in chromosome segregation can occur at meiosis I, when homologues generally separate, or at meiosis II, when sister chromatids separate. Because recombination occurs in meiosis I before disjunction but does not occur at centromeres, errors during meiosis I result in chromosomes with different (homologous) haplotypes at their centromeres, whereas sister chromatids nondisjoined in meiosis II have the same (sister) haplotype at their centromeres (Fig. 3a). (Sex chromosomes X and Y disjoin in meiosis I, and the sister chromatids of X and Y disjoin at meiosis II.) Encouragingly, for chromosome 21—the principal chromosome for which earlier estimates were possible—our finding of 33% meiosis I events and 67% meiosis II events matched previous estimates from trisomy 21 patients with paternal-origin gains[36].

Across all chromosomes, meiosis I gains and meiosis II gains had very different relative frequencies in different individuals and on different chromosomes (Fig. 3d, e). For example, sex chromosomes were 2.2 times more likely to be affected in meiosis I than meiosis II, whereas autosomes were 2.0 times more likely to be affected in meiosis II than meiosis I (proportion test two-sided $P = 1.3 \times 10^{-6}$). The lack of correlation between meiosis I and meiosis II vulnerabilities (Fig. 3d, e) indicated that meiosis I and II are differentially challenging to different chromosomes and to different people.

Although crossovers are required for proper chromosomal segregation[37] and seem to be protective against nondisjunction in maternal meiosis, in which chromosomes are maintained in diplotene of meiosis I for decades[8], the relationship of crossovers to aneuploidy is less clear in paternal meiosis[24,36,38–41]. We found that chromosome gains originating in meiosis I—when recombination occurs—had 36% fewer total crossovers than matched, well-segregated chromosomes did (Supplementary Methods), suggesting that crossovers protected against meiosis I nondisjunction of the chromosomes on which they occurred (Extended Data Fig. 9b and Supplementary Notes). No similar

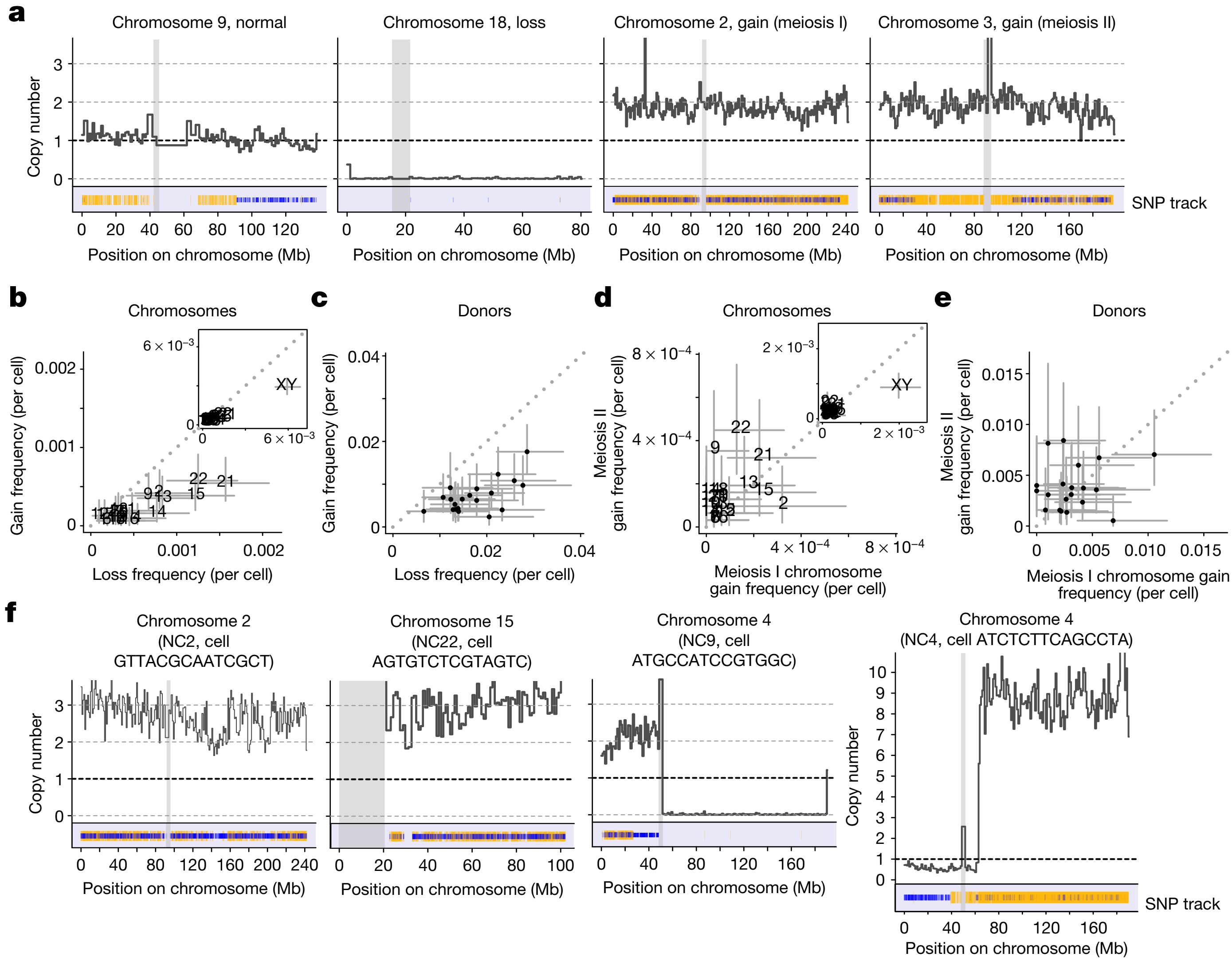


Fig. 3 | Aneuploidy in sperm from 20 sperm donors. a, Example chromosomal ploidy analyses. Thick dark grey line, DNA copy number measurement (normalized sequence coverage in 1-Mb bins); blue (haplotype 1) and yellow (haplotype 2) vertical lines, observed heterozygous SNP alleles, plotted with 90% transparency; grey vertical boxes, centromeres (based on the reference hg38 human genome). **b**–**e**, Frequencies (number of events divided by number of cells) of various aneuploidy categories. **b**, **d**, $n = 23$ chromosomes; **c**, **e**, $n = 20$ donors. Error bars are 95% binomial confidence intervals. **b**, Frequencies of whole-chromosome losses versus gains for each chromosome (excluding XY chromosomes, Pearson's $r = 0.88$, two-sided $P = 7 \times 10^{-8}$; including XY chromosomes (inset), Pearson's $r = 0.99$, two-sided $P < 10^{-300}$). **c**, Per-sperm-donor aneuploidy rates (excluding XY (not shown), Pearson's $r = 0.51$, two-sided $P = 0.02$; including XY, Pearson's $r = 0.62$, two-sided $P = 0.003$). **d**, Frequencies of whole-chromosome gains occurring during meiosis I versus meiosis II for each chromosome (excluding XY, Pearson's $r = 0.32$, two-sided $P = 0.15$; including XY (inset), Pearson's $r = 0.85$, two-sided $P = 3 \times 10^{-7}$). **e**, Frequencies of whole-chromosome gains occurring during meiosis I versus II for each donor (excluding XY (not shown), Pearson's $r = 0.06$, two-sided $P = 0.80$; including XY, Pearson's $r = 0.17$, two-sided $P = 0.47$). **f**, Example genomic anomalies detected in sperm cells, plotted as in **a**. NC2, NC4, NC9 and NC22 signify individual sperm donors; cells are identified by 14-bp DNA barcode sequences.

relationship was observed for meiosis II gains (although the simulated control distribution for meiosis II is inherently less accurate; Supplementary Notes) or at other levels of aggregation (Extended Data Fig. 9b–d and Supplementary Notes).

Other chromosome-scale genomic anomalies

Many sperm had complex patterns of aneuploidy that could not be explained by the canonical single-chromosome missegregation event. We detected 19 gametes that had three, instead of one, copies of entire or nearly entire chromosomes (2, 15, 20 and 21; Fig. 3f and Extended Data Fig. 10a, b). Chromosome 15 was particularly likely to be present in two extra copies; in fact, sperm with three copies of all or most of chromosome 15 ($n = 10$) outnumbered sperm with two copies of chromosome 15 ($n = 2$) (Fisher's exact test versus Poisson two-sided $P = 2 \times 10^{-7}$) (Supplementary Notes).

Other gametes carried anomalies encompassing incomplete chromosomes. These included: one cell that gained the *p* arm of chromosome 4 while losing the *q* arm; cells with gains of two copies of a chromosome arm; and cells with losses of chromosome arms (Fig. 3f and Extended Data Fig. 10c, d). One cell carried at least eight copies of most of the *q* arm of chromosome 4 (Fig. 3f). This gamete—which we estimate contained almost a billion base pairs of extra DNA—carried both parental haplotypes of chromosome 4, though almost all of the roughly eight copies came from just one of the parental haplotypes (93% of observed alleles in the amplified region were haplotype 2). It is likely that diverse mutational processes generate these genomic anomalies (Supplementary Discussion).

Discussion

Interindividual variation in crossover rates has previously been inferred from SNP data from families[2–7,9–12]. Here, highly parallel single-gamete sequencing has revealed that sperm donors with high crossover rates also exhibit closer crossover spacing, even when controlling for the number of crossovers actually made on a chromosome. On the basis of these analyses, we consider it most likely that interindividual variation in crossover interference is the true driver of variation in crossover rate and placement.

These same constellations of correlated meiotic crossover phenotypes—low interference, high rates and use of centromere-proximal zones—tended to characterize the same gametes from any donor. Cells with more crossovers in half of their genome tended to have more crossovers in the other half, to have made consecutive pairs of crossovers closer together in genomic distance—even when making just two crossovers on a chromosome—and to have placed proportionally more of their crossovers in nondistal chromosomal regions.

We considered what could cause these meiotic phenotypes to covary across chromosomes, in individual cells, and among people. The physical length of chromosomes during meiosis, which reflects their compaction, has been observed to vary up to twofold among individual spermatocytes while being strongly correlated across chromosomes in the same spermatocyte; spermatocytes with more-compacted chromosomes also generally have fewer incipient crossovers[20,21,42]. A unifying model (Extended Data Fig. 11) explains the covariance of these meiotic phenotypes while providing a candidate mechanism for interindividual variation: cell-to-cell variation in the compaction of meiotic chromosomes—and person-to-person variation in the average degree of this compaction—would cause these phenotypes to covary in the manner observed in Fig. 2b–e.

Our enthusiasm for this model relies on several additional earlier observations (Extended Data Fig. 11). First, at a cellular level, crossover interference occurs as a function of physical (micrometre-scale) distance along the meiotic chromosome axis or synaptonemal complex, rather than as a function of genomic (base-pair) distance[43–45]. Second, the first crossover on a chromosome is more likely to occur distally[13–15]. Such a model also predicts a shared mechanism for sex differences in recombination rates and interindividual variation among individuals of the same sex: oocytes have a longer synaptonemal complex, more crossovers and decreased crossover interference (as measured in genomic distances) than spermatocytes, but have the same synaptonemal-complex length extent of crossover interference[22,42,46,47].

Human genetics research has revealed that recombination phenotypes are heritable and associate with common variants at many genomic loci[3–6,11,12]. A recent genome-wide association study found that variation in crossover rate and placement is associated with variants near genes that encode components of the synaptonemal complex, which connects and compacts meiotic chromosomes, and with genes involved in the looping of homologues along the chromosome axis[3]. Our model predicts that inherited genetic variation at these loci may bias the average degree of compaction of meiotic chromosomes; the fact that this same property varies among cells from the same donor[20,21] shows that variance is well-tolerated and compatible with diverse-but-successful meiotic outcomes.

The sharing of covarying phenotypes between the single-cell and person-to-person levels suggests that a core biological mechanism shapes both inter- and intra-individual (single-cell) variation in meiotic outcomes. Such parallelisms between cell-biological and human-biological variation could in principle exist in a wide variety of biological contexts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2347-0.

1. Broman, K. W. & Weber, J. L. Characterization of human crossover interference. *Am. J. Hum. Genet*. **66**, 1911–1926 (2000).
2. Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).
3. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
4. Kong, A. et al. A high-resolution recombination map of the human genome. *Nat. Genet*. **31**, 241–247 (2002).
5. Kong, A. et al. Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet*. **46**, 11–16 (2014).
6. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
7. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
8. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat. Rev. Genet*. **13**, 493–504 (2012).
9. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet*. **63**, 861–869 (1998).
10. Cheung, V. G., Burdick, J. T., Hirschmann, D. & Morley, M. Polymorphic variation in human meiotic recombination. *Am. J. Hum. Genet*. **80**, 526–530 (2007).
11. Chowdhury, R., Bois, P. R., Feingold, E., Sherman, S. L. & Cheung, V. G. Genetic analysis of variation in human meiotic recombination. *PLoS Genet*. **5**, e1000648 (2009).
12. Fledel-Alon, A. et al. Variation in human recombination rates and its genetic determinants. *PLoS ONE* **6**, e20321 (2011).
13. Brown, P. W. et al. Meiotic synapsis proceeds from a limited number of subtelomeric sites in the human male. *Am. J. Hum. Genet*. **77**, 556–566 (2005).
14. Gruhn, J. R. et al. Correlations between synaptic initiation and meiotic recombination: a study of humans and mice. *Am. J. Hum. Genet*. **98**, 102–115 (2016).
15. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A. & Hassold, T. Cytological studies of human meiosis: sex-specific differences in recombination originate at, or prior to, establishment of double-strand breaks. *PLoS ONE* **8**, e85075 (2013).
16. Baudat, F. & de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res*. **15**, 565–577 (2007).
17. Plug, A. W., Xu, J., Reddy, G., Golub, E. I. & Ashley, T. Presynaptic association of Rad51 protein with selected sites in meiotic chromatin. *Proc. Natl Acad. Sci. USA* **93**, 5920–5924 (1996).
18. Ioannou, D., Fortun, J. & Tempest, H. G. Meiotic nondisjunction and sperm aneuploidy in humans. *Reproduction* **157**, R15–R31 (2018).
19. Templado, C., Uroz, L. & Estop, A. New insights on the origin and relevance of aneuploidy in human spermatozoa. *Mol. Hum. Reprod*. **19**, 634–643 (2013).
20. Lynn, A. et al. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* **296**, 2222–2225 (2002).
21. Wang, S. et al. Per-nucleus crossover covariation and implications for evolution. *Cell* **177**, 326–338 (2019).
22. Hou, Y. et al. Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).
23. Kirkness, E. F. et al. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res*. **23**, 826–832 (2013).
24. Lu, S. et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).
25. Ottolini, C. S. et al. Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nat. Genet*. **47**, 727–735 (2015).
26. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
27. Miller, D., Brinkworth, M. & Iles, D. Paternal DNA packaging in spermatozoa: more than the sum of its parts? DNA, histones, protamines and epigenetics. *Reproduction* **139**, 287–301 (2010).
28. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
29. Zheng, G. X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol*. **34**, 303–311 (2016).
30. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D. & Auton, A. Escape from crossover interference increases with maternal age. *Nat. Commun*. **6**, 6260 (2015).
31. Berg, I. L. et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet*. **42**, 859–863 (2010).
32. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet*. **40**, 1124–1129 (2008).
33. Hinch, A. G. et al. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**, eaau8861 (2019).
34. Housworth, E. A. & Stahl, F. W. Crossover interference in humans. *Am. J. Hum. Genet*. **73**, 188–197 (2003).
35. Sun, F. et al. Human male recombination maps for individual chromosomes. *Am. J. Hum. Genet*. **74**, 521–531 (2004).
36. Oliver, T. R. et al. Investigation of factors associated with paternal nondisjunction of chromosome 21. *Am. J. Med. Genet. A*. **149A**, 1685–1690 (2009).

37. Page, S. L. & Hawley, R. S. Chromosome choreography: the meiotic ballet. *Science* **301**, 785–789 (2003).
38. Sun, F. et al. The relationship between meiotic recombination in human spermatocytes and aneuploidy in sperm. *Hum. Reprod*. **23**, 1691–1697 (2008).
39. Ferguson, K. A., Wong, E. C., Chow, V., Nigro, M. & Ma, S. Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. *Hum. Mol. Genet*. **16**, 2870–2879 (2007).
40. Ma, S., Ferguson, K. A., Arsovska, S., Moens, P. & Chow, V. Reduced recombination associated with the production of aneuploid sperm in an infertile man: a case report. *Hum. Reprod*. **21**, 980–985 (2006).
41. Savage, A. R. et al. Elucidating the mechanisms of paternal non-disjunction of chromosome 21 in humans. *Hum. Mol. Genet*. **7**, 1221–1227 (1998)
42. Tease, C. & Hultén, M. A. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet. Genome Res*. **107**, 208–215 (2004).
43. Wang, S., Zickler, D., Kleckner, N. & Zhang, L. Meiotic crossover patterns: obligatory crossover, interference and homeostasis in a single process. *Cell Cycle* **14**, 305–314 (2015).
44. Zhang, L., Liang, Z., Hutchinson, J. & Kleckner, N. Crossover patterning by the beam-film model: analysis and implications. *PLoS Genet*. **10**, e1004042 (2014).
45. Zhang, L. et al. Topoisomerase II mediates meiotic crossover interference. *Nature* **511**, 551–556 (2014).
46. Billings, T. et al. Patterns of recombination activity on mouse chromosome 11 revealed by high resolution mapping. *PLoS ONE* **5**, e15340 (2010).
47. Petkov, P. M., Broman, K. W., Szatkiewicz, J. P. & Paigen, K. Crossover interference underlies sex differences in recombination rates. *Trends Genet*. **23**, 539–542 (2007).

Methods

A companion protocol for generating single-sperm libraries using the methods presented here is available via Protocol Exchange[48]. Custom scripts (available via Zenodo[49]) are referenced by name in the Methods sections describing the relevant analyses. Recombination and aneuploidy data generated by the methods described are also publicly available[50]. All statistical analyses were performed in *R* unless otherwise noted. Details of further analysis methods are provided in the Supplementary Methods.

Sample information

Sperm samples from 20 anonymous, karyotypically normal sperm donors were obtained from New England Cryogenic Center under a 'not human subjects' determination from the Harvard Faculty of Medicine Office of Human Research Administration (protocols M23743-101 and IRB16-0834). Donors consented at the time of initial donation for samples to be used for research purposes. The 'not human subjects' determination was based on the use of discarded biospecimens for which research consent had been obtained, and on the fact that researchers had no interactions with the biospecimen donors and no access to identifiable information about the biospecimens. The reviewing committee also reviewed and approved our deposition of the data into a National Institutes of Health (NIH) repository. All experiments were performed in accordance with all relevant guidelines and regulations. (Specimens can be obtained from the New England Cryogenic Center upon Institutional Review Board (IRB) approval.) No statistical methods were used to predetermine sample size. As no conditions or experimental groups were analysed for this study, no randomization or blinding was performed.

Samples arrived in liquid nitrogen in 'egg yolk buffer' or 'standard buffer with glycerol' (no further buffer information provided), and were aliquoted and stored in liquid nitrogen in the same buffers.

Per sperm-bank policy, donors are 18–38 years old at the time of donation and the precise age of donors is not released. Donor identifiers used here were created specifically for this study and are not linked to any external identifiers.

ddPCR to evaluate genome accessibility

To evaluate how often regions from two different chromosomes co-occurred (as would be expected from cells), we performed droplet digital polymerase chain reaction (ddPCR) with naked DNA, untreated sperm cells or sperm cells decondensed as described below but with variable heat incubation times. For each assay targeting each chromosome, we created a 20× assay mix by combining 25.2 µl of 100 µM forward primer (from IDT), 25.2 µl of 100 µM reverse primer (IDT) and 7 µl of 100 µM probe (IDT for fluorescein amidite (FAM)-labelled probes; Life Technologies for VIC-labelled probes) with 82.6 µl ultrapure water. We carried out ddPCR as previously described[51], following section 3.2 steps 4–12, but with untreated sperm or sperm DNA florets as input instead of DNA.

For this analysis, we targeted chromosome 7 with an assay directed to intergenic region chr7:106552149–106552176 (hg38): forward primer sequence CGTAATGGGGCACAGGGATATA; reverse primer sequence CTGTGAGAGGTAGAGAATCGCC; probe sequence CAC AGAGTCCATTTGCAGCACCTCAGT; probe fluorophore FAM. We targeted chromosome 10 with an assay for the *RPP30* gene at chr10:92631759–92631820: forward primer sequence GATTTGGA CCTGCGAGCG; reverse primer sequence GCGGCTGTCTCCACAAGT; probe sequence CTGACCTGAAGGCTCT; probe fluorophore VIC. We calculated the percentage of molecules expected to be linked from each reaction as previously described[52].

Sperm cell library generation

We generated accessible sperm nuclei 'florets' using a combination of published decondensation protocols[53,54] with some modifications. Sperm aliquots containing more than 200,000 cells were thawed on ice and then washed by spinning for 10 min at 400*g* at 4 °C. The pellet was resuspended in 10 µl phosphate-buffered saline (PBS, Gibco/LifeTechnologies) and recentrifuged under the same conditions. The sperm pellet was resuspended in 2.5 µl of a sucrose buffer containing 250 mM sucrose (Sigma), 5 mM $MgCl_2$ (Sigma) and 10 mM Tris HCl (pH 7.5, Thermo Scientific). Sperm aliquots were submerged in liquid nitrogen and immediately quick-thawed by holding them in a warm fist; three such freeze–thaw cycles were performed.

Freeze-thawed sperm solution was combined with 22.5 µl decondensation buffer (113 mM KCl (Sigma), 12.5 mM KH_2PO_4 (Sigma), 2.5 mM Na_2HPO_4 (Sigma), 2.5 mM $MgCl_2$ (Sigma) and 20 mM Tris (Thermo Scientific) freshly supplemented with 150 µM heparin (sodium salt from porcine, Sigma catalogue number H3393) and 2 mM β-mercaptoethanol (Sigma)). The reaction was incubated at 37 °C for 45 min. To allow enzymatic DNA amplification, heparin was inactivated by mixing the sperm solution with 0.5 U heparinase I (Sigma H2519) by gently pipetting and incubating at room temperature for 2 h (ref. [55]).

The sperm solution was moved to ice, and sperm floret concentration was determined by diluting 1:100 with PBS and staining with 1× SYBR I (Thermo Scientific), then counting using the green fluorescence channel at 10× magnification.

Droplets were prepared using the following modifications to 10× Genomics' GemCode (version 1; ref. [29]) user guide revision C (in place of steps 5.1–5.3.9); all reagents come from the 10× Genomics GemCode kit. Ultrapure water was combined with 10,833 sperm to a final volume of 5 µl; 10,000 sperm were used for library generation. To each sperm sample was added 60 µl of a master mix containing 32.5 µl GemCode reagent mix, 1.5 µl primer release agent, 9.2 µl GemCode polymerase and 16.8 µl ultrapure water.

GemCode beads were vortexed at full speed for 25 s, and then diluted 1:11 with ultrapure water to a total volume of at least 90 µl per sample. Per 10× Genomics' GemCode's protocol, 60 µl of the sample-master mix combination was added to the droplet generation chip, followed by 85 µl of freshly pipette-mixed 1:11-diluted bead mixture and 150 µl of droplet generation oil.

Droplets were generated and processed by library generation following 10× Genomics' GemCode (version 1) user guide revision C (step 5.3.10 through to the end of section 6).

Sequencing and sequence data processing

We generated two libraries per sperm donor and additional libraries for four initial samples with low cell counts. We sequenced four or five libraries at a time on S2 200 cycle flow cells on an Illumina NovaSeq. The read structure was 178 cycles for read 1, 8 cycles for read 2 (index read one), 14 cycles for read 3 (index read two containing the cell barcode; later treated as the reverse read), and 5 cycles for read 4 (unused; included to fulfil the NovaSeq's paired-end requirement).

To convert the data to mapped binary alignment map (BAM) files with cell and molecular barcodes encoded as read tags, we used Picard Tools v.2.2 (http://broadinstitute.github.io/picard) and Drop-seq Tools v.2.2 (https://github.com/broadinstitute/Drop-seq/releases; see https://github.com/broadinstitute/Drop-seq/blob/master/doc/Drop-seq_Alignment_Cookbook.pdf for details on running many of the tools)[28].

Illumina binary base call (BCL) files were converted to unmapped BAM files using Picard's ExtractIlluminaBarcodes and IlluminaBasecallsToSam with read structure 178T8B14T (cell barcodes, present in the i5 index, were incorporated as read 2 for ease of downstream processing). BAMs were processed to include unique molecular identifiers (UMIs) and cell barcodes as read tags, and to exclude reads with poor-quality cell barcodes or UMIs; consequently, each read was retained as single-end with a 14-base-pair (bp) cell barcode stored in tag XC and a 10-bp molecular barcode/UMI stored in tag XM. The first 10 bp of read 1 were used as the UMI. First, DropSeq

Tools' TagBamWithReadSequenceExtended was called with BASE_RANGE = 1-14, BASE_QUALITY = 10, BARCODED_READ = 2, DISCARD_READ = true, TAG_NAME = XC, NUM_BASES_BELOW_QUALITY = 1. Subsequently, TagBamWithReadSequenceExtended was called again with BASE_RANGE = 1-10, BASE_QUALITY = 10, HARD_CLIP_BASES = true, BARCODED_READ = 1, DISCARD_READ = false, TAG_NAME = XM, NUM_BASES_BELOW_QUALITY = 1. Finally, DropSeq Tools' FilterBAM was called with parameter TAG_REJECT = XQ.

Reads were aligned to hg38 using bwa mem[56] v.0.7.7-r441. BAMs were converted to FastQ using Picard's SamToFastQ, FastQ reads were aligned using bwa mem −M, and then unmapped BAMs were merged with mapped BAMs using Picard's MergeBamAlignment, with non-default options INCLUDE_SECONDARY_ALIGNMENTS = false and PAIRED_RUN = false. Reads were marked PCR duplicates using Drop-seq Tools' SpermSeqMarkDuplicates (part of Drop-seq tools v2.2 and above) with options STRATEGY = READ_POSITION, CELL_BARCODE_TAG = XC, MOLECULAR_BARCODE_TAG = XM, NUM_BARCODES = 20000, CREATE_INDEX = true. BAM files for all lanes and index sequences from the same sample were merged using Picard's MergeSamFiles before alignment and/or during duplicate marking with all BAMs given as input to SpermSeqMarkDuplicates.

Variant calling and sperm cell genotyping

For each donor, we pooled all reads from all libraries, including reads that did not derive from a barcode associated with a complete sperm cell. Using GATK v.3.7 (refs. [57,58]) in hg38, we followed GATK's best-practices documentation for base quality score recalibration; for genomic variable call format (gVCF) generation using HaplotypeCaller (in 'discovery' mode with −stand_call_conf 20); and for joint genotyping with GenotypeGVCFs. We filtered variants with SelectVariants −selectType SNP and VariantFiltration (−filterExpression 'QD < 3.0'). We then performed variant quality score recalibration (VQSR) following GATK's best practices, except that we excluded annotations MQ and DP (VariantRecalibrator with GATK provided resources; −an QD, MQRankSum, ReadPosRankSum, FS and SOR; −mode SNP; −trustAllPolymorphic; and tranches 90, 99.0, 99.5, 99.9 and 100.0). We applied tranche 99.9 recalibration using ApplyRecalibration −mode SNP and obtained the names of SNPs from SNP database (dbSNP) build 146 (ref. [59]) using VariantAnnotator −dbsnp. We filtered our sites to contain only those biallelic SNPs that were present in Hardy–Weinberg equilibrium in 1000 Genomes Phase 3 (ref. [60]) using SelectVariants −concordance with a VCF containing only these sites (from GATK's resource bundle). We excluded SNPs in centromeric regions or acrocentric arms, as defined by the University of California Santa Cruz (UCSC)'s Genome Browser's cytoband track[61,62] (http://genome.ucsc.edu; the same centromere boundaries were used in all analyses), and those in known paralogous regions[63]. We selected only heterozygous SNPs using SelectVariants −selectType SNP−selectTypeToExclude INDEL−restrictAllelesTo BIALLELIC−excludeFiltered−setFilteredGtToNocall−selectexpressions 'vc.getGenotype("'"<sample name>"'").isHet()'.

We identified the SNPs present in each sperm cell and the allele that was present using GenotypeSperm (part of Drop-seq Tools v.2.2 and above). For downstream analyses, we generated a file with columns cell, pos and gt, with gt having the value 0 for the reference allele and 1 for the alternate allele for SNPs that had one or more UMIs covering only one base matching the reference or alternate allele (see our script gtypesperm2cellsbyrow.R).

Chromosome-scale phasing

We identified barcodes that were potentially associated with cells by plotting the cumulative fraction of reads associated with each ranked barcode and identifying the inflection point of this curve (Extended Data Fig. 1f). We then included only those barcodes that had substantial read depth on either the X or the Y chromosome but not both, as the vast majority of sperm cells should contain only one sex chromosome. (We later added these barcodes back in before formally identifying and excluding cell doublets.)

To phase sperm donors' genomes, we used all quality-controlled heterozygous sites in these cell barcodes expected to correspond to sperm cells, excluding observations of SNPs for which the observed allele was not the reference or alternate allele in the parental genome, or for which more than one allele was observed. For each chromosome, we converted per-cell SNP calls into 'fragments' for input into the HapCUT phasing software[64,65] by considering each consecutive pair of SNPs observed in a cell to be a fragment (see our script gtypesperm2fmf.R). We then used HapCUT with parameter −maxiter 100 to generate chromosomal phase. After identifying and removing cell doublets (see below), we repeated phasing with only non-doublet cell barcodes.

To validate our phasing method, we simulated single-cell SNP observations from known haplotypes, including 2% genotype errors and a variable percentage of cell doublets. In brief, sites were randomly sampled from one known haplotype of chromosome 17 until a crossover location was probabilistically assigned on the basis of the deCODE recombination map[6], then sampled from the other haplotype (one crossover was simulated per cell). To simulate PCR or sequencing errors, 2% of the sites were randomly assigned to an allele. Doublets were simulated by combining two cells and retaining 70% of the observed sites at random. We performed five random simulations for each doublet proportion, for the mean proportion of sites 'observed' in each cell, and for the number of cells simulated, and then followed our phasing protocol using each simulation (see our script simulatespermseqfromhaps.py).

To further validate phasing, we used Sperm-seq data to phase one donor's genome and compared these phased haplotypes to this donor's Eagle[66,67]-generated haplotypes. We compared the phase relationship between each consecutive pair of SNPs (identifying the proportion of switch errors between the two phased sets). We also compared the Sperm-seq allele–allele phase of all pairs of alleles in perfect linkage disequilibrium in 1000 Genomes Phase 3 (ref. [60]) in the populations matching the donor's ancestry.

Cell doublets

To identify cell barcodes associated with more than one sperm cell (cell doublets), we detected consecutively observed SNP alleles that appeared on different parental haplotypes, which could occur because of crossover, error, or the presence of two haplotypes in the same droplet (doublet). We ranked barcodes by the proportion of consecutive SNPs that spanned haplotypes by using all SNPs from all autosomes except the autosome with the most haplotype-spanning consecutive SNPs (so as to avoid mistakenly identifying cells with chromosome gains as doublets); this resulted in a clear inflection point wherein cell doublets had a quickly accelerating proportion of haplotype-spanning consecutive SNPs (Extended Data Fig. 2d–f). All cell barcodes below this inflection point (identified with the function 'ede' from the R package 'inflection' https://CRAN.R-project.org/package=inflection) were considered non-doublet (Extended Data Fig. 2f) (see our script computeSwitchesandInflThresh.R). Even though we specifically exclude the autosome with the most haplotype-spanning consecutive SNPs from doublet identification, any cells with multiple chromosome gains (especially more than two) or whole-genome diploidy would be excluded by this method.

Crossover events

We identified crossover events on all autosomes (but excluded the *p* arms of acrocentric chromosomes for which SNPs were excluded from analysis) by finding transitions between tracts of SNPs with alleles that match different parental haplotypes using a hidden Markov model written in R with package 'HMM' (https://CRAN.R-project.org/package=HMM). To ensure that we detected crossovers located near the ends of SNP coverage (subtelomeric regions are frequently used for crossovers in spermatogenesis), we ran the HMM in both the forward-chromosomal and the reverse-chromosomal directions, with the start probability for one haplotype equal to 1 if the first two SNPs observed were of that haplotype. In addition to two states for parental haplotypes, we included a third 'error' state to capture cases in which a haplotype 1 allele is observed in a haplotype 2 region (and vice versa), for example, owing to PCR or sequencing error, gene conversion, or cases in which a small piece of off-haplotype ambient DNA was captured in a droplet. Crossovers were where one haplotype transitioned to another, or where one haplotype transitioned to the error state and then to the other haplotype. Crossover boundaries were the last SNP in the first haplotype and the first in the next. The key parameters for this algorithm are the transition probability between haplotypes (set to 0.001, from the per-cell median 26 crossovers divided by the per-cell median 24,710 heterozygous SNPs) and transition probability into and out of the 'error' state (we set the probability of transition into this state to 0.03 from either haplotype, as only a few percent of SNPs are off-haplotype; we set the probability of staying in error to 0.9 to allow for the occasional tract of SNPs from an ambient piece of off-haplotype DNA). Emission probabilities were 100% haplotype 1 alleles from haplotype 1, 100% haplotype 2 alleles from haplotype 2, and equal probability haplotype 1 or 2 alleles from the third 'error' state. Crossover calling was robust to a range of low transition probabilities (see our script spseqHMMCOCaller_3state.R, which calls crossovers on one chromosome).

After aneuploidy identification, we marked aneuploid chromosomes as having no crossovers for all crossover analyses (absent chromosomes have no crossovers and crossovers are called differently on gained chromosomes, described below).

Identifying even-coverage cell barcodes

We used Genome STRiP v2.0 (http://software.broadinstitute.org/software/genomestrip/)[68,69] to determine sequence read depth (observed number of reads divided by expected number of reads) in bins of 100 kb of uniquely mappable sequence across the genome in each sperm cell, using Genome STRiP's default guanine–cytosine (GC) bias correction and repetitive region masking for reference genome gr38. We divided the read depth by two to obtain the read depth per haploid rather than diploid genome. Input to Genome STRiP was a BAM file containing only cells of interest, with read groups set to < sample name>:<cell barcode> (created using Drop-seq Tools' ConvertTagToReadGroup with options CELL_BARCODE_TAG = XC, SAMPLE_NAME = <name of sample/donor>, CREATE_INDEX = true, and CELL_BC_FILE = list of barcodes potentially associated with cells, described above).

A minority of cell barcodes were associated with eccentric read depth across many chromosomes, with wave-like read depth vacillating between 0 and 2 or more. (We hypothesize that these cell barcodes were associated with sperm nuclei that did not properly decondense, such that some regions of the genome were more accessible than others, leading to undulating read depths across more- and less-accessible chromatin.) To identify and exclude such barcodes, we treated read depths across each chromosome as a time series and used Box–Jenkins autoregressive integrated moving average (ARIMA) modelling to model how read depth observations relied on their previous values and their overall averages (implemented via the R package 'forecast'[70,71], excluding differencing). By visual inspection, we determined that chromosomes with certain ARIMA criteria were likely to have an undulating read depth, and that cell barcodes with five or more such identified chromosomes were likely to have eccentric read depths globally. We flagged individual chromosomes if: (1) the sum of the AR1 and AR2 coefficients was greater than 0.7, the AR1 coefficient was greater than 0.9, or the net sum of all AR and MA coefficients was greater than 1.25; and (2) either the net sum of AR and MA coefficients was greater than 0.4 or the intercept was less than 0.8 or greater than 1.2. If both criteria in (2) were met, this signified an exceedingly odd chromosome, which we counted twice. Cell barcodes with five or more chromosomes flagged in this way were excluded from downstream analyses. (Because gains of large amounts of the genome cause artificially depressed read depths on nongained chromosomes, we manually examined any cells with a large range of ARIMA intercepts and more than five chromosomes denoted as unstable. Any such cells that had simply gained a large proportion of the genome—for example, three copies of chromosome 2—were included rather than excluded). We cross-referenced all cell exclusions with called aneuploidies, confirming that cells were not excluded simply on the basis of having lost or gained a chromosome (see our scripts setupgsreaddepth.R, exclbadreaddepth_arima_1.R, exclbadreaddepth_initid_2.R, and exclbadreaddepth_finalize_3.R).

Replicate barcodes ('bead doublets')

One sperm cell can be encapsulated in a droplet with more than one barcoded bead. To identify such cases, where pairs of sperm genomes were identical, we determined the proportion of SNPs that were of the same haplotype for each pair of barcodes. We imputed the haplotype of all heterozygous SNPs on the basis of the haplotype of surrounding observed SNPs and locations of recombination events, and compared SNP haplotypes across sperm cell pairs. SNP observations between boundaries of crossovers were excluded from analysis. Sperm cells shared on average 50% of their genomes, but a few sets of barcodes shared nearly 100% of their SNP haplotypes (Extended Data Fig. 3a). We considered these pairs to be 'bead doublets' or replicate barcodes. In all downstream analyses, we used only one barcode (chosen randomly) from a set corresponding to the same cell (see our scripts imputeHaplotypeAllSNPs.R, compareSpermHapsPropSNPs.R, combineChrsSpermHapsPropSNPs.R, and curateNonRepBCList.R).

Crossover zones

To define regions of recombination use, we found local minima of the density (built-in function in R) of all crossovers' median positions across all samples on each chromosome. Minima were identified using the findPeaks function (from https://github.com/stas-g/findPeaks) on the inverse density with $m = 3$. Crossover zones run from the beginning of the chromosome (including the whole *p* arm for acrocentric chromosomes) to the location of the first local minimum, from the location of the first local minimum plus one base pair to the next local minimum, and so on, with the last zone on each chromosome ending at the chromosome end (see our script findcozones_peaks.R).

Aneuploidy and chromosome arm loss/gain

As described previously (see Methods section 'Identifying even-coverage cell barcodes'), we used Genome STRiP (http://software.broadinstitute.org/software/genomestrip/)[68,69] to determine read depth in each sperm cell in 100-kb bins. We located chromosomes or chromosome arms with aberrant read depth to identify aneuploidy.

We excluded genomic regions that had outlying read depths across all cells, defined as those with $P < 0.05$ in a one-sided one-sample *t*-test (looking for increased read depth) against the expected mean read depth of 2# (defined below). To identify gains of autosomes, we performed a one-sided one-sample *t*-test (expecting increased read depth in a gain) for each cell against the expected read depth for a gain of one copy, 2#. For each cell, this analysis compared the distribution of all bins' read depth across a region of interest to the gain expectation 2#, and flagged any cells whose read depth distributions were not significantly different ($P \geq 0.05$). We used the same approach to identify losses, comparing a cell's read depth distribution across bins to 0.1 and flagging any that were not significantly higher ($P \geq 0.05$).

The expected copy number for gains is 2, but the expected read depth for gains depends on the size of the chromosome: a library corresponding to a cell with a chromosome gain has more reads than would be in that same library without a gain. This phenomenon pulls the read depth down globally by increasing the total number of expected reads, causing the denominator in each read depth bin (the expected number of reads in that bin) to increase. Therefore, we computed a chromosome-specific critical read depth value for identifying gains: 2# = 2*(the proportion of the genome in base pairs coming from all chromosomes other than the tested one). For losses, we used 0.1 rather than 0 as the expected read depth, because a small number of reads generally align to every chromosome in every library.

For nonacrocentric chromosomes, we performed aneuploidy calling for the arms separately and for the whole chromosome. Because amplification of more than two copies of a chromosome arm could result in the whole chromosome passing the *P*-value threshold, we required a whole-chromosome event to pass the *P*-value threshold at the whole-chromosome level and to have a rounded read depth for both arms of 2 or more for a gain (or 0 for a loss). For the acrocentric chromosomes, only the *q* arm was considered, and any *q* arm gain or loss was considered to be a whole-chromosome event (unless investigated further).

For the sex chromosomes, we followed a similar statistical framework, but a loss was considered an aneuploidy only if both the X and the Y chromosomes were flagged as lost. A gain was called if both X and Y chromosomes were present (see our scripts setupgsreaddepth.R, idaneus_initialttests.R, curateaneudata_clean.R, getautosomalaneumatrix.R and getxykaryos_aneus.R for aneuploidy calling and output formatting; see our scripts curateAnFreqFromCodeMatrix.R, curateInitAnalyzeXYKaryos.R and combineAnFreq_AutXY.R for conversion of outputs of aneuploidy calling to cross-donor aneuploidy frequency tables).

Division of origin for chromosome gains

To see when chromosome gains originated, we determined whether the centromeres of the multiple copies of the chromosomes were heterozygous and therefore from homologues, which typically disjoin in meiosis I, or homozygous and therefore from sister chromatids, which typically disjoin in meiosis II. We identified heterozygous regions for all cells using a hidden Markov model (HMM) in which the states are: (1) heterozygous (emitting either haplotype's alleles), or (2) homozygous (emitting only one haplotype's alleles), with the transition probability between the states equal to the recombination transition probability. For each gain, we determined whether heterozygous tracts overlapped the centromere. If a heterozygous tract started before the start of the centromere and ended after the end of the centromere, or started at the first SNP observed on an acrocentric chromosome or within the first ten SNPs and was more than ten SNPs long, then chromosome was classified as a meiosis I gain; if no heterozygous tract overlapped the centromere, it was classified as a meiosis II gain (see our scripts getDiploidTracts_hmm.R, originOfGainID.R and curateOriginMultSamps.R).

At the sex chromosomes, any XY sex chromosome gain derives from meiosis (X and Y are homologues), whereas an XX or YY gain derives from meiosis II (sister chromatids are duplicated).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Crossover and aneuploidy data (individual events and counts per donor and/or per cell), including the source data underlying Figs. 2, 3b–e and Extended Data Figs. 5–9, are available via Zenodo at https://doi.org/10.5281/zenodo.2581570. Raw sequence data are available in the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) via the Database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/) for general research use upon application and approval (study accession number phs001887.v1.p1).

Code Availability

Analysis scripts and documentation are available via Zenodo at https://doi.org/10.5281/zenodo.2581595.

48. Bell, A. D., Mello, C. J. & McCarroll, S. A. Sperm-seq wet lab protocol: sperm preparation and droplet-based sequencing library generation. *Protoc. Exch*. https://doi.org/10.21203/rs.3.pex-823/v1 (2020).
49. Bell, A. D. et al. Analysis scripts for: Insights about variation in meiosis from 31,228 human sperm genomes. Zenodo, https://doi.org/10.5281/zenodo.2581595 (2019).
50. Bell, A. D. et al. Recombination and aneuploidy data for: Insights about variation in meiosis from 31,228 human sperm genomes. Zenodo, https://doi.org/10.5281/zenodo.2581570 (2019).
51. Bell, A. D., Usher, C. L. & McCarroll, S. A. Analyzing copy number variation with droplet digital PCR. *Methods Mol. Biol*. **1768**, 143–160 (2018).
52. Regan, J. F. et al. A rapid molecular approach for chromosomal phasing. *PLoS ONE* **10**, e0118270 (2015).
53. Montag, M., Tok, V., Liow, S. L., Bongso, A. & Ng, S. C. In vitro decondensation of mammalian sperm and subsequent formation of pronuclei-like structures for micromanipulation. *Mol. Reprod. Dev*. **33**, 338–346 (1992).
54. Samocha-Bone, D. et al. In-vitro human spermatozoa nuclear decondensation assessed by flow cytometry. *Mol. Hum. Reprod*. **4**, 133–137 (1998).
55. Taylor, A. C. Titration of heparinase for removal of the PCR-inhibitory effect of heparin in DNA samples. *Mol. Ecol*. **6**, 383–385 (1997).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
57. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).
58. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
59. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. **29**, 308–311 (2001).
60. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
61. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res*. **12**, 996–1006 (2002).
62. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res*. **45** (D1), D626–D634 (2017).
63. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med*. **371**, 2477–2487 (2014).
64. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
65. Selvaraj, S., Dixon, R. J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol*. **31**, 1111–1118 (2013).
66. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet*. **48**, 1443–1448 (2016).
67. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet*. **48**, 811–816 (2016).
68. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet*. **43**, 269–276 (2011).
69. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet*. **47**, 296–303 (2015).
70. Hyndman, R. J. & Khandakar, Y. Automatic time series forecasting: the forecast package for R. *J. Stat. Softw*. **26**, 1–22 (2008).
71. Hyndman, R. et al. forecast: forecasting functions for time series and linear models. R package version 8.4, http://pkg.robjhyndman.com/forecast (2018).
72. Kauppi, L. et al. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* **331**, 916–920 (2011).
73. Kleckner, N., Storlazzi, A. & Zickler, D. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends Genet*. **19**, 623–628 (2003).
74. Revenkova, E. et al. Cohesin SMC1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. *Nat. Cell Biol*. **6**, 555–562 (2004).
75. Zickler, D. & Kleckner, N. Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet*. **33**, 603–754 (1999).
76. Wang, S. et al. Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. *Cell* **168**, 977–989 (2017).
77. Blat, Y., Protacio, R. U., Hunter, N. & Kleckner, N. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell* **111**, 791–802 (2002).

Acknowledgements We thank G. Genovese for suggestions on analyses; E. Macosko for advice on technology development; other members of the McCarroll laboratory, including C. Whelan, S. Burger and B. Handsaker, for advice; M. Daly, J. Hirschhorn, S. Elledge and S. Schilit for their insights; 10× Genomics for discussions about reagents; C. L. Usher and C. K. Patil for contributions to the text and figures; and those who commented on the preprint version of this article for their input. This work was supported by grant R01 HG006855 to S.A.M.; a Broad Institute NextGen award to S.A.M.; and a Harvard Medical School Program in Genetics and Genomics NIH Ruth L. Kirchstein training grant to A.D.B.

Author contributions A.D.B. and S.A.M. conceived and led the studies. A.D.B, S.A.M. and C.J.M. developed the experimental methods. A.D.B. and C.J.M. performed all experiments, generating all data. A.D.B and S.A.M. designed the strategies for crossover and aneuploidy analysis, and A.D.B. performed the crossover and aneuploidy analyses. A.D.B., J.N. and A.W. wrote the sequence and variant processing software, pipelines and analytical methods. A.D.B. wrote the software for crossover calling and analysis. A.D.B. and S.A.B wrote the software for aneuploidy calling. A.D.B. and S.A.M. wrote the manuscript with contributions from all authors.

Competing interests A.D.B. and S.A.M. are inventors on a United States Provisional Patent application (PCT/US2019/029427; applicant: President and Fellows of Harvard College), currently in the PCT stage, relating to droplet-based genomic DNA capture, amplification and sequencing that is capable of obtaining high-throughput single-cell sequence from individual mammalian cells, including sperm cells. A.D.B. was an occasional consultant from Ohana Biosciences between October 2019 and March 2020. The other authors declare no competing interests.

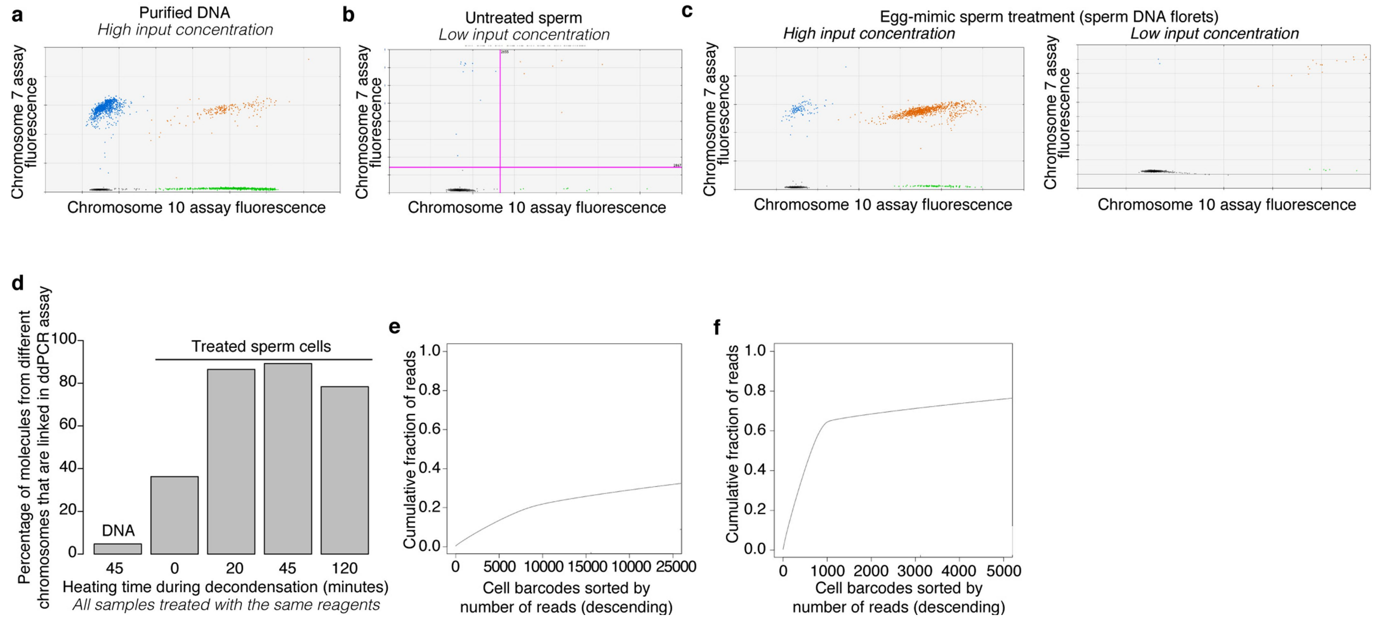
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2347-0.

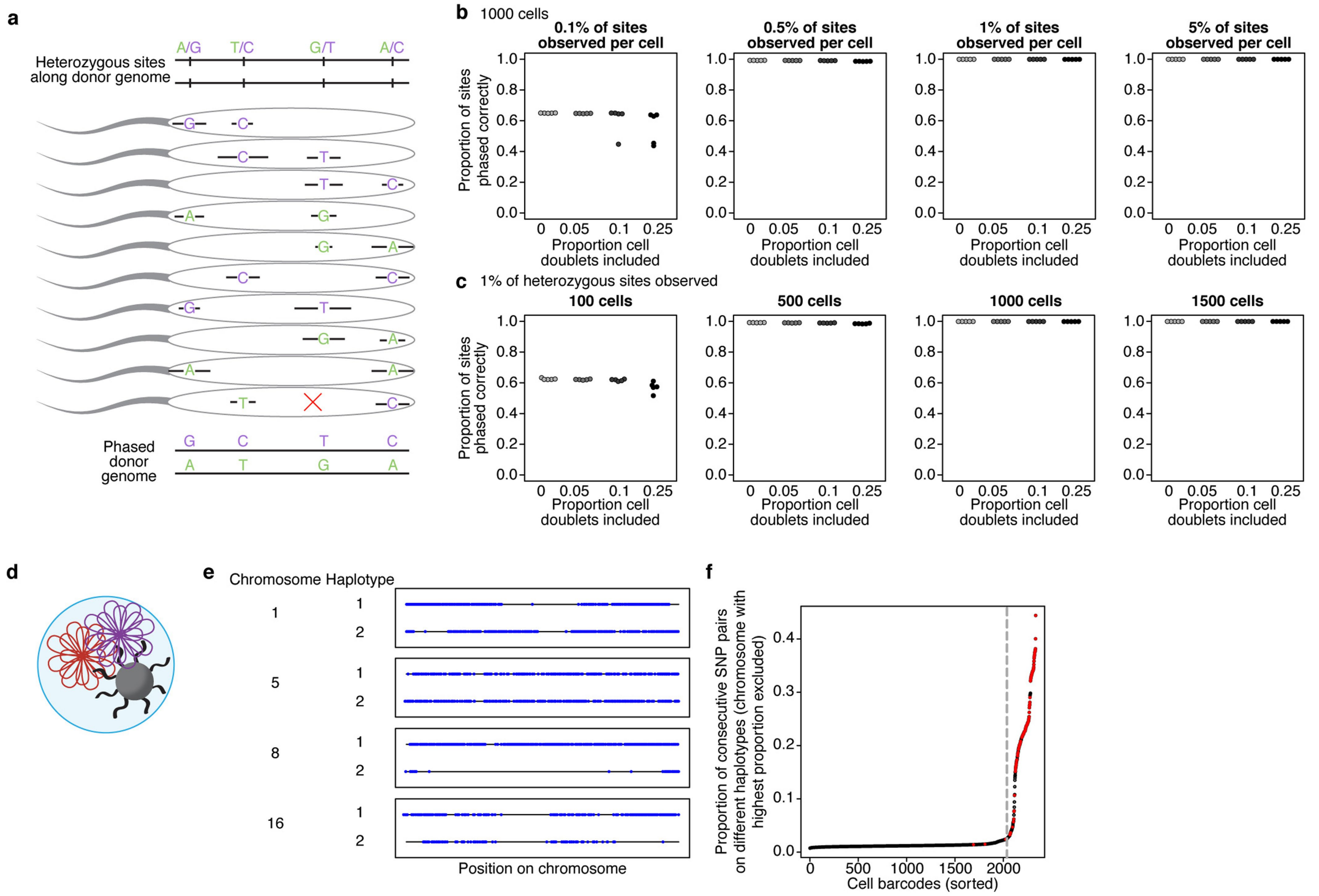
Correspondence and requests for materials should be addressed to A.D.B. or S.A.M.

Peer review information *Nature* thanks Donald Conrad, Beth Dumont, Augustine Kong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

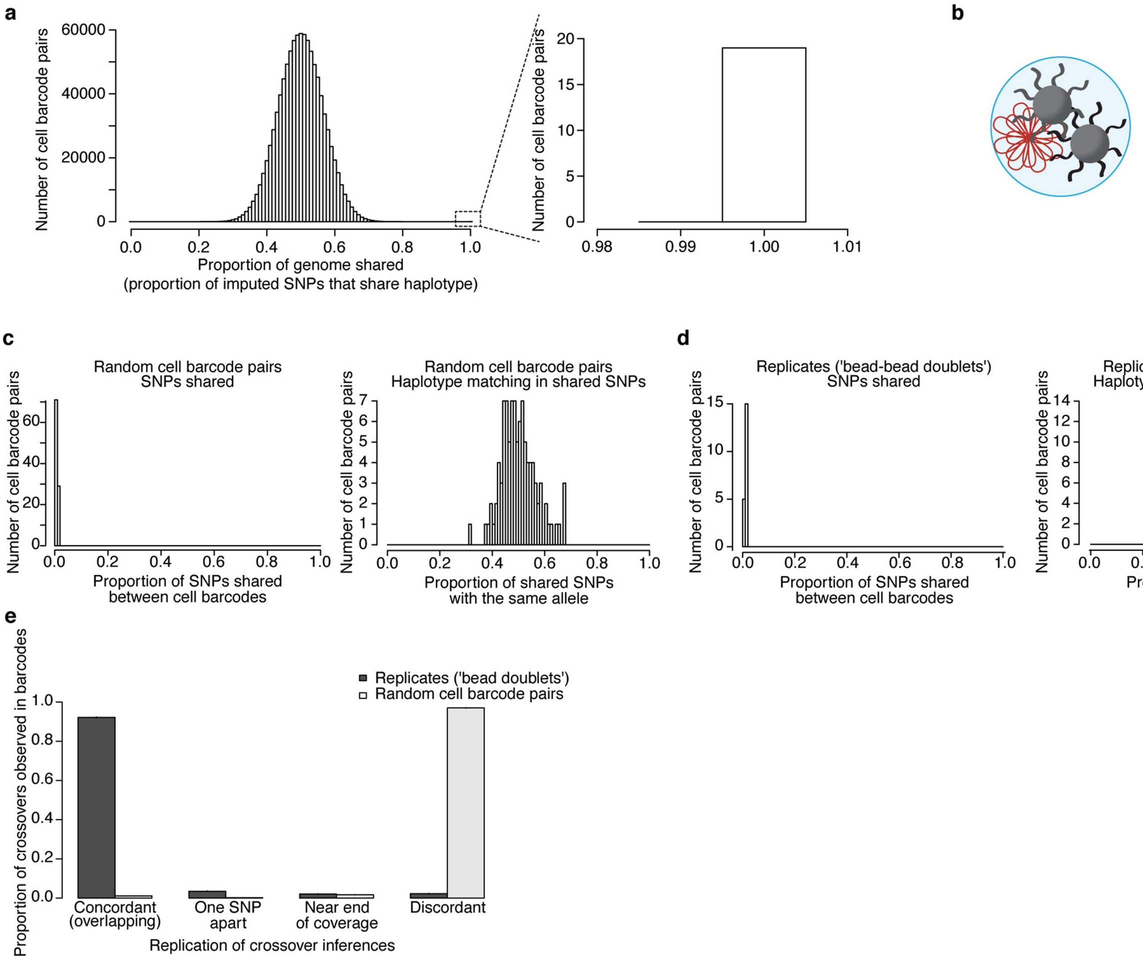
Reprints and permissions information is available at http://www.nature.com/reprints.



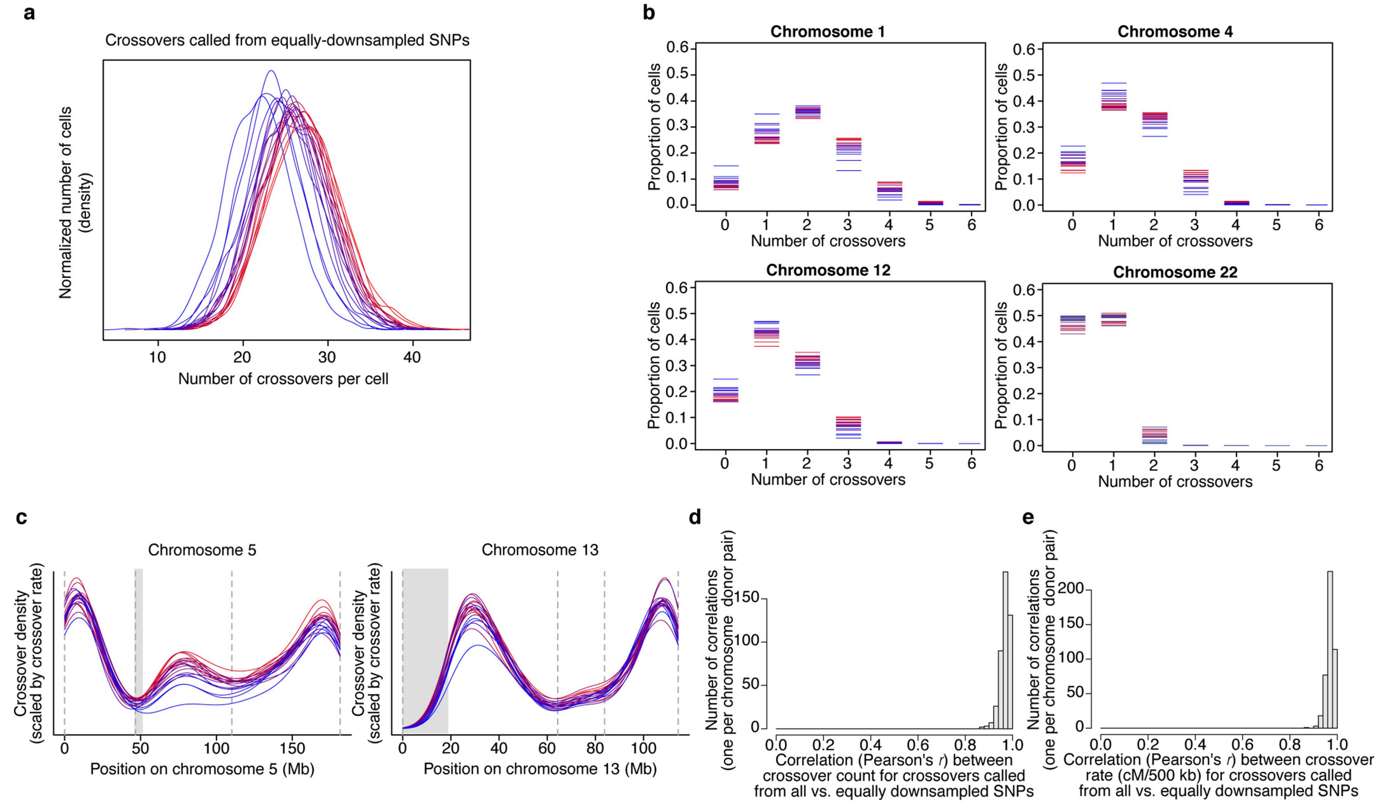
Extended Data Fig. 1 | Characterization of egg-mimic sperm preparation and optimization of bead-based single-sperm sequencing. a–c, Two-channel fluorescence plots showing the results of ddPCR with the input template noted above each panel, demonstrating that two loci (from different chromosomes) are detectable in the same droplet far more often when sperm DNA florets (rather than purified DNA) are used as input. Each point represents one droplet. Grey points (bottom left) represent droplets in which neither template molecule was detected; blue points (top left) represent droplets in which the assay detected a template molecule for the locus on chromosome 7; green droplets (bottom right) represent droplets in which the assay detected a template molecule for the locus on chromosome 10; and brown points (top right) represent droplets in which both loci were detected. With a high concentration of purified DNA as input (**a**), comparatively fewer droplets contain both loci than when untreated (**b**) or treated (**c**) sperm were used as input. Sperm 'florets' treated with the egg-mimicking decondensation protocol had a much higher fraction of droplets containing both loci than did purified DNA (compare **a** with **c**, left) and had more-sensitive ascertainment and cleaner results (quadrant separation) than untreated sperm (compare **b** with **c**, right). The pink lines in **b** delineate the boundaries between droplets categorized as negative or positive for each assay. **d**, Optimization of sperm preparation: characterization of the effect of different lengths of 37 °C incubation of sperm cells treated with egg-mimicking decondensation reagents on how often the loci on chromosomes 7 and 10 were detected in the same ddPCR droplet. The *y*-axis shows the percentage of molecules that are calculated to be linked to each other (that is, physically linked in input) for assays targeting chromosomes 7 and 10. Extracted DNA ('DNA', a negative control) gives the expected result of random assortment of the two template molecules into droplets. The 45-min heat treatment was used for all subsequent experiments in this study. **e**, **f**, Distribution of sequence reads across cell barcodes from droplet-based single-sperm sequencing. Each panel shows the cumulative fraction of all reads from a sequencing run coming from each read-number-ranked cell barcode; a sharp inflection point delineates the barcodes with many reads from those with few reads. Points to the left of the inflection point are the cell barcodes that are associated with many reads (that is, beads that are coencapsulated with cells); the height of the inflection point reflects the proportion of the sequence reads that come from these barcodes. Only reads that mapped to the human genome (hg38) and were not PCR duplicates are included. **e**, Data from an initial adaptation of 10× Genomics' GemCode linked reads system[29], where a small proportion of the reads come from cell barcodes associated with putative cells. **f**, Data from the final, implemented adaptation of 10× Genomics' GemCode linked reads system[29] for the same number of input sperm nuclei as in **e**. The *x*-axis in **f** includes five times fewer barcodes than in **e**.



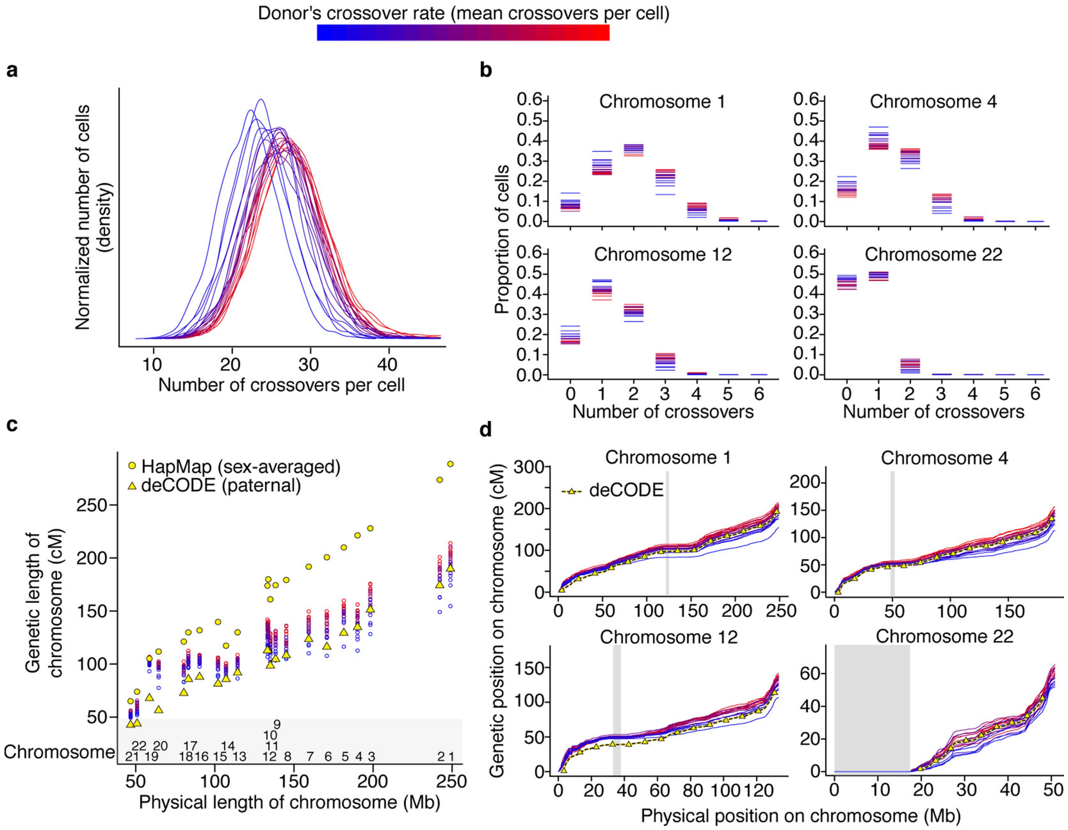
Extended Data Fig. 2 | Evaluation of chromosomal phasing and identification of cell doublets. **a**, Phasing strategy. Green and purple denote the chromosomal phase of each allele (unknown before analysis). Each sperm cell carries one parental haplotype (green or purple), except where a recombination event separates consecutively observed SNPs (red X in bottom sperm). Because alleles from the same haplotype will tend to be observed in the same sperm cells, the haplotype arrangement of the alleles can be assembled at whole-chromosome scale (resulting in the phased donor genome). **b**, Evaluation of our phasing method using 1,000 simulated single-sperm genomes (generated from two a priori known parental haplotypes and sampled at various levels of coverage). Because cell doublets (which combine two haploid genomes and potentially two haplotypes at any region) can in principle undermine phasing inference, we included these doublets in the simulation (in proportions shown on the *x*-axis, which bracket the observed doublet rates). Each point shows the proportion of SNPs phased concordantly with the correct (a priori known) haplotypes (*y*-axis) for one simulation (five simulations were performed for each unique combination of proportion of cell doublets and percentage of sites observed). **c**, Relationship of phasing capability to number of cells analysed. Data are as in **b**, but for different numbers of simulated cells. All simulations had an among-cell mean of 1% of heterozygous sites observed. **d**, A cell doublet: when two cells (here, sperm DNA florets) are encapsulated together in the same droplet, their genomic sequences will be tagged with the same barcode; such events must be recognized computationally and excluded from downstream analyses. **e**, Four example chromosomes from a cell barcode associated with two sperm cells (a cell doublet). Black lines show haplotypes; blue circles are observations of alleles, shown on the haplotype from which they derive. Both parental haplotypes are present across regions of chromosomes for which the cells inherited different haplotypes. **f**, Computational recognition of cell doublets in Sperm-seq data (from an individual sperm donor, NC11). We used the proportion of consecutively observed SNP alleles derived from different parental haplotypes to identify cell doublets; this proportion is generally small (arising from sparse crossovers, PCR/sequencing errors, and/or ambient DNA) but is much higher when the analysed sequence comes from a mixture of two distinct haploid genomes. We use 21 of the 22 autosomes to calculate this proportion, excluding the autosome with the highest such proportion (given the possibility that a chromosome is aneuploid). The dashed grey line marks the inflection point beyond which sperm genomes are flagged as potential doublets and excluded from downstream analysis. Red points indicate barcodes with coverage of both X and Y chromosomes (potentially X + Y cell doublets or XY aneuploid cells); black points indicate barcodes with one sex chromosome detected (X or Y). The red (XY) cells below the doublet threshold are XY aneuploid but appear to have just one copy of each autosome.



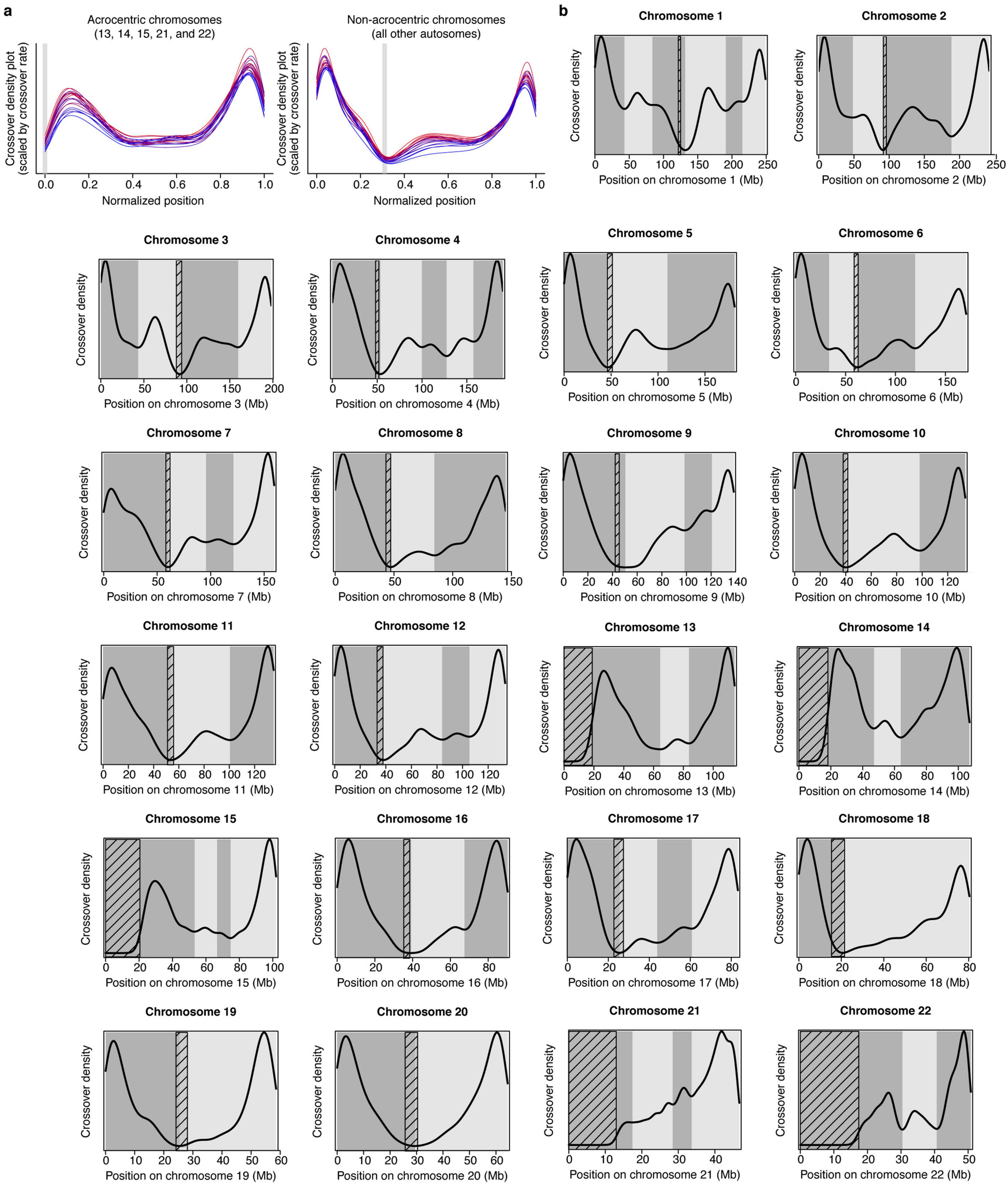
Extended Data Fig. 3 | Identification and use of 'bead doublets'. a, SNP alleles were inferred genome-wide (for each sperm genome) by imputation from the subset of alleles detected in each cell and by Sperm-seq-inferred parental haplotypes. For each pair of sperm genomes (cell barcodes), we estimated the proportion of all SNPs at which they shared the same imputed allele. A small but surprising number of such pairwise comparisons (19 of 984,906 from the donor shown, NC14) indicates essentially identical genomes (ascertained through different SNPs). **b**, We hypothesize that this arises from a heretofore undescribed scenario that we call 'bead doublets', in which two barcoded beads have coencapsulated with the same gamete and whose barcodes therefore tagged the same haploid genome. **c**, Random pairs of cell barcodes (here 100 pairs selected from donor NC10) tend to investigate few of the same SNPs (left), and also tend to detect the same parental haplotype on average at the expected 50% of the genome (right). **d**, 'Bead doublet' barcode pairs (here 20 pairs from donor NC10, who had the median number of bead doublets, left) also investigate few of the same SNPs, yet detect identical haplotypes throughout the genome (right). Results were consistent across donors. **e**, Use of 'bead doublets' to characterize the concordance of crossover inferences between distinct samplings of the same haploid genome by different barcodes. The bead doublets (barcode pairs) were compared to 100 random barcode pairs per donor. Crossover inferences were classified as 'concordant' (overlapping, detected in both barcodes), as 'one SNP apart' (separated by just one SNP, detected in both barcodes), as 'near end of coverage' (within 15 heterozygous SNPs of the end of SNP coverage at a telomere, where the power to infer crossovers is partial), or as discordant. Error bars (with small magnitude) show binomial 95% confidence intervals for the number of crossovers per category divided by number of crossovers total in both barcodes (32,714 crossovers total in 1,201 bead doublet pairs; 67,862 crossovers total in 2,000 random barcode pairs; some barcodes are in multiple bead doublets or random barcode pairs).



Extended Data Fig. 4 | Numbers and locations of crossovers called from downsampled data (equal number of SNPs in each cell, randomly chosen). To eliminate any potential effect of unequal sequence coverage across donors and cells, we used downsampling to create datasets with equal coverage (numbers) of heterozygous SNP observations in each cell. Crossovers were called from these random equally sized sets of SNPs from all cells. **a**, **b**, Crossover number per cell globally (**a**) and per chromosome (**b**) (785,476 total autosomal crossovers called from downsampled SNPs included, 30,778 cells included, aneuploid chromosomes excluded). **c**, Density plots of crossover location with crossover midpoints plotted and area scaled to be equal to the per-chromosome crossover rate. Grey rectangles mark centromeric regions; coordinates are in hg38. **d**, Similar numbers of crossovers were called from full data and equally downsampled SNP data: we performed correlation tests across cells for each donor and chromosome to compare the number of crossovers called from all data to the number of crossovers called from equal numbers of randomly downsampled SNPs. The histogram shows Pearson's *r* values for all 460 (20 donors × 23 chromosomes (total number plus number for 22 autosomes)) tests (*n* per test = 974–2,274 cells per donor as in Extended Data Table 1; all chromosome comparisons Pearson's $r > 0.83$; all two-sided $P < 10^{-300}$). **e**, Crossovers called from equally downsampled SNP data were in similar locations to those called from all data: we performed correlation tests comparing crossover rates in 500-kb bins (centimorgans (cM) per 500 kb) from all data versus equally downsampled SNP data for each donor and chromosome. The histogram shows Pearson's *r* values for all 460 (20 donors × 23 chromosomes (genome-wide rate plus rate for 22 autosomes)) tests (*n* per test = number of 500-kb bins per chromosome (genome-wide: 5,739; chromosomes 1–22: 497, 484, 396, 380, 363, 341, 318, 290, 276, 267, 270, 266, 228, 214, 203, 180, 166, 160, 117, 128, 93, 101); all chromosome comparisons Pearson's $r > 0.87$, all two-sided $P < 10^{-300}$).

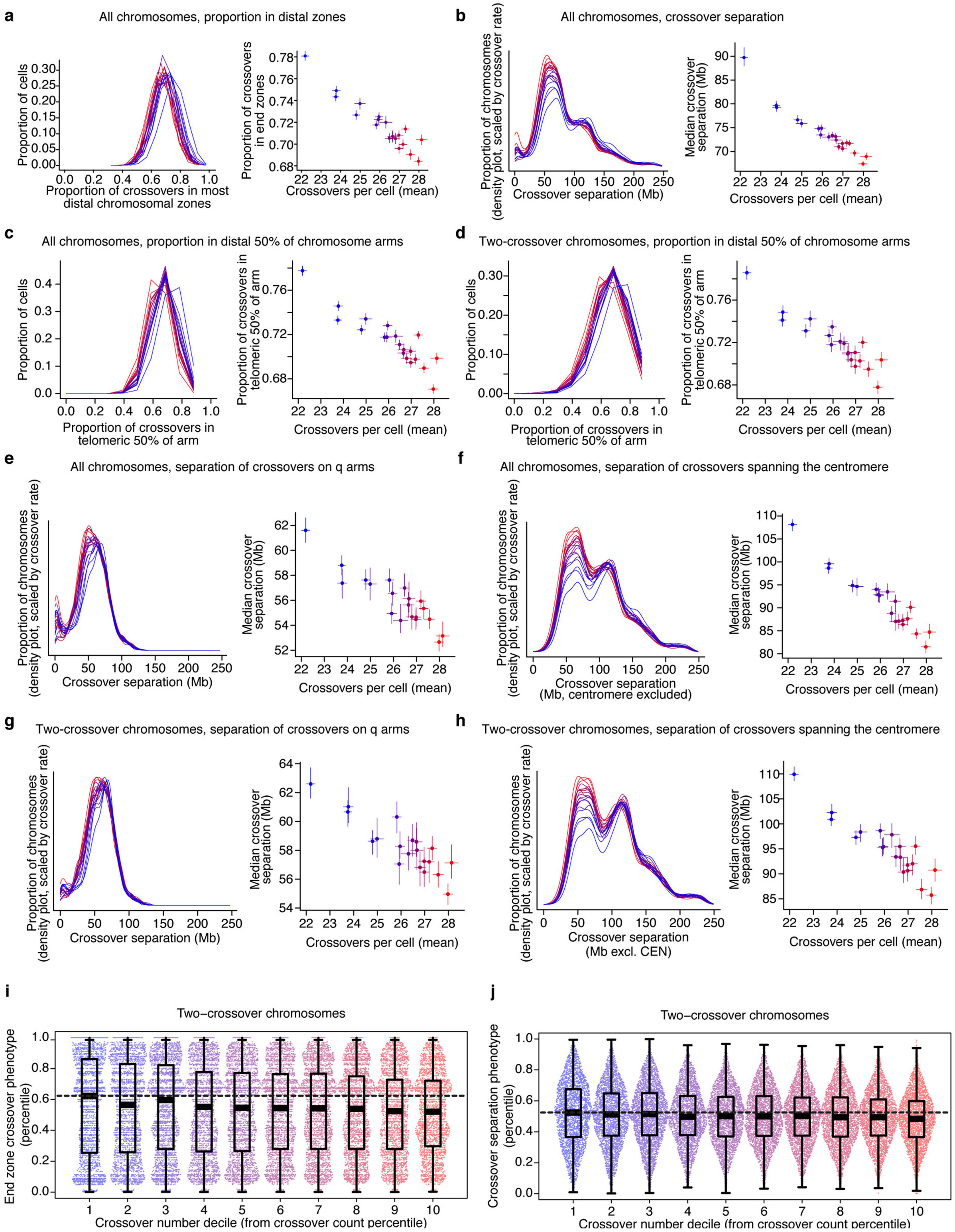


Extended Data Fig. 5 | Interindividual and intercell recombination rate from single-sperm sequencing. a, Density plot showing the per-cell number of autosomal crossovers for all 31,228 cells (813,122 total autosomal crossovers) from 20 sperm donors (per-donor cell and crossover numbers as in Extended Data Table 1; aneuploid chromosomes were excluded from crossover analysis). Colours represent a donor's mean crossover rate (crossovers per cell) from low (blue) to high (red). This same mean recombination rate derived colour scheme is used for donors in all figures. The recombination rate differs among donors ($n = 20$; Kruskal–Wallis chi-squared = 3,665; df = 19; $P < 10^{-300}$). **b**, Per-chromosome crossover number in each of the 20 sperm donors (data as in **a** but shown for individual chromosomes). **c**, Per-chromosome genetic map lengths for: each of the 20 sperm donors, as inferred from Sperm-seq data (colours from blue to red reflect donors' individual crossover rates as in **a**); a male average, as estimated from pedigrees by deCODE[6] (yellow triangles); and a population average (including female meioses, which have more crossovers), as estimated from HapMap data[7] (yellow circles). The deCODE genetic maps stop 2.5 Mb from the ends of SNP coverage. **d**, Physical versus genetic distances (for individualized sperm donor genetic maps and deCODE's paternal genetic map) plotted at 500-kb intervals (in hg38 coordinates). Grey boxes denote centromeric regions (or centromeres and acrocentric arms). Sperm-seq maps are broadly concordant with deCODE maps (see the correlation test results in Supplementary Notes), except at subtelomeric regions that are not included in deCODE's map.



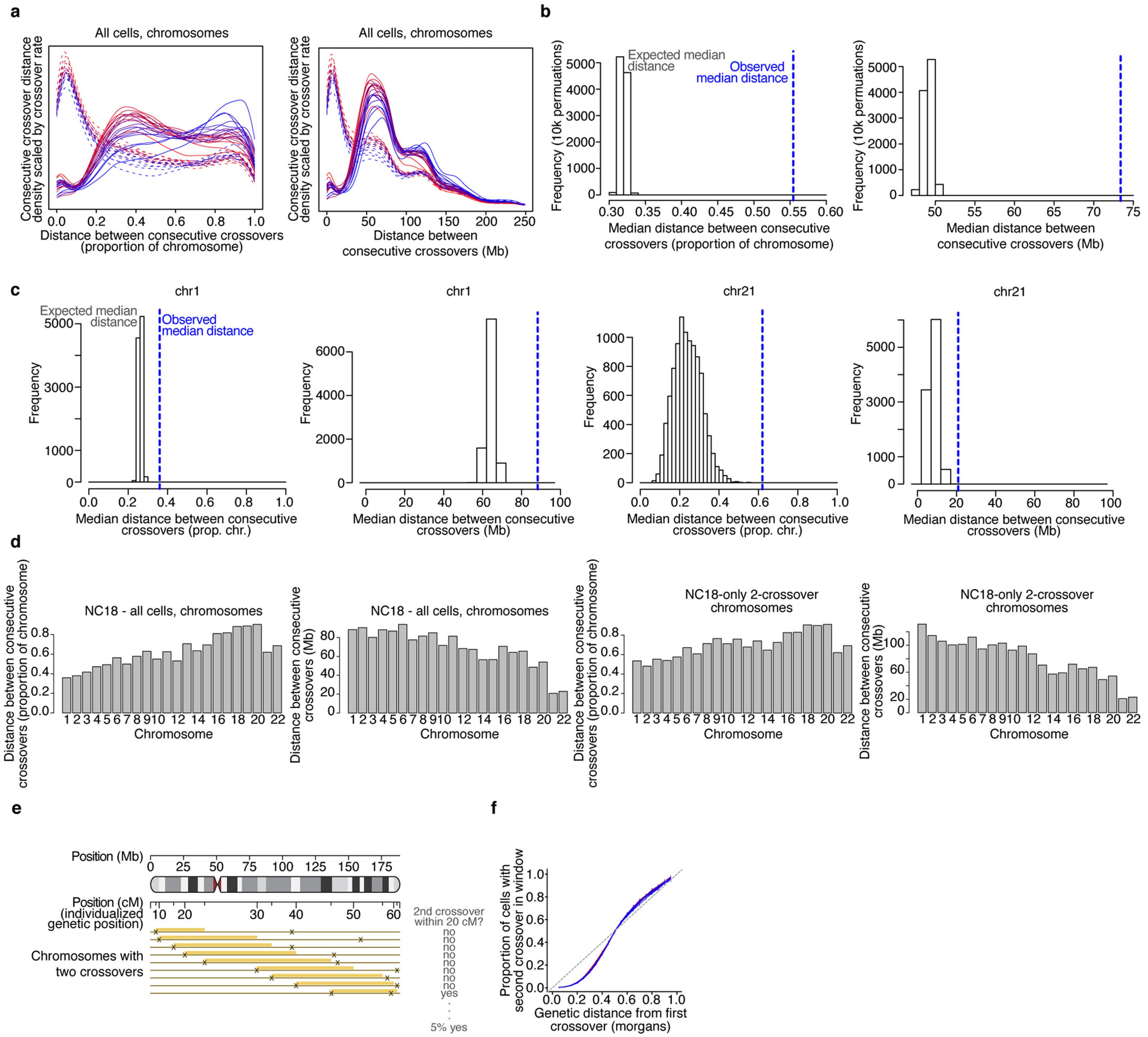
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Distributions of crossover locations along chromosomes (in 'crossover zones'). a, Each donor's crossover locations are plotted as a coloured line; the colour indicates the donor's overall crossover rate (blue, low; red, high); grey boxes show the locations of centromeres (or, for acrocentric chromosomes, of centromeres and *p* arms). We used the midpoint between the SNPs bounding each inferred crossover as the position for each crossover in all analyses. To combine data across chromosomes, we show crossover locations (density plot) on 'meta-chromosomes' in which crossover locations are normalized to the length of the chromosome or arm on which they occurred. For acrocentric chromosomes, only the *q* arm was considered; for nonacrocentric chromosomes, the *p* and *q* arms were afforded space on the basis of the proportion of the nonacrocentric genome (in base pairs) they comprise, with the centromere placed at the summed *p* arms' proportion of base pairs of these chromosomes. Crossover locations were first converted to the proportion of the arm at which they fall, and then these positions were normalized to the genome-wide *p* or *q* arm proportion. **b**, Identification of chromosomal zones of recombination use ('crossover zones') from all donors' crossovers for 22 autosomes. Density plots are shown of crossover location for all sperm donors' total 813,122 crossovers (aneuploid chromosomes excluded; the crossover location is the midpoint between SNPs bounding crossovers) along autosomes (hg38). Crossover zones (bounded by local minima of crossover density) are shown with alternating shades of grey. Diagonally hatched rectangles indicate centromeres (or centromeres and acrocentric arms).

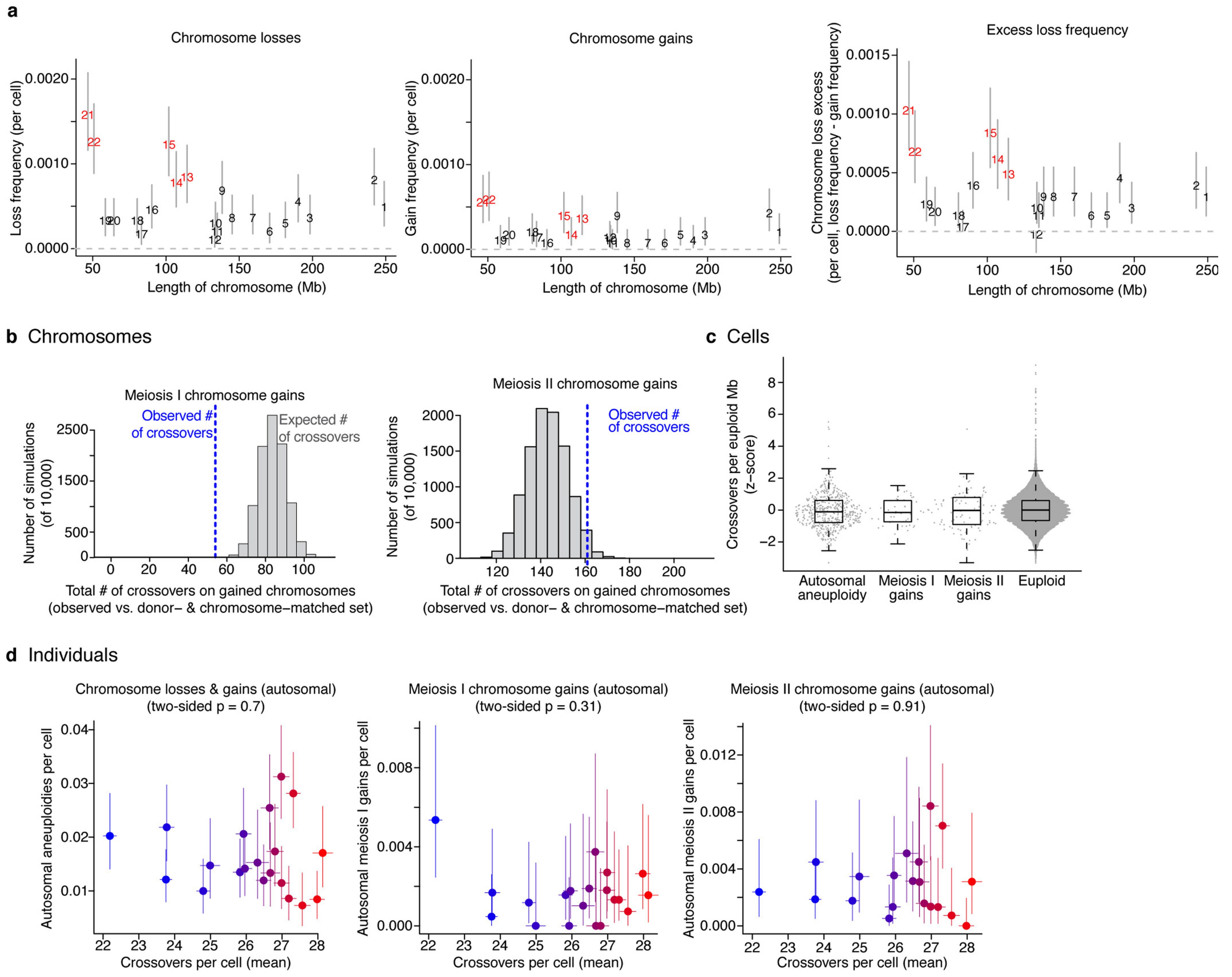


Extended Data Fig. 7 | See next page for caption.

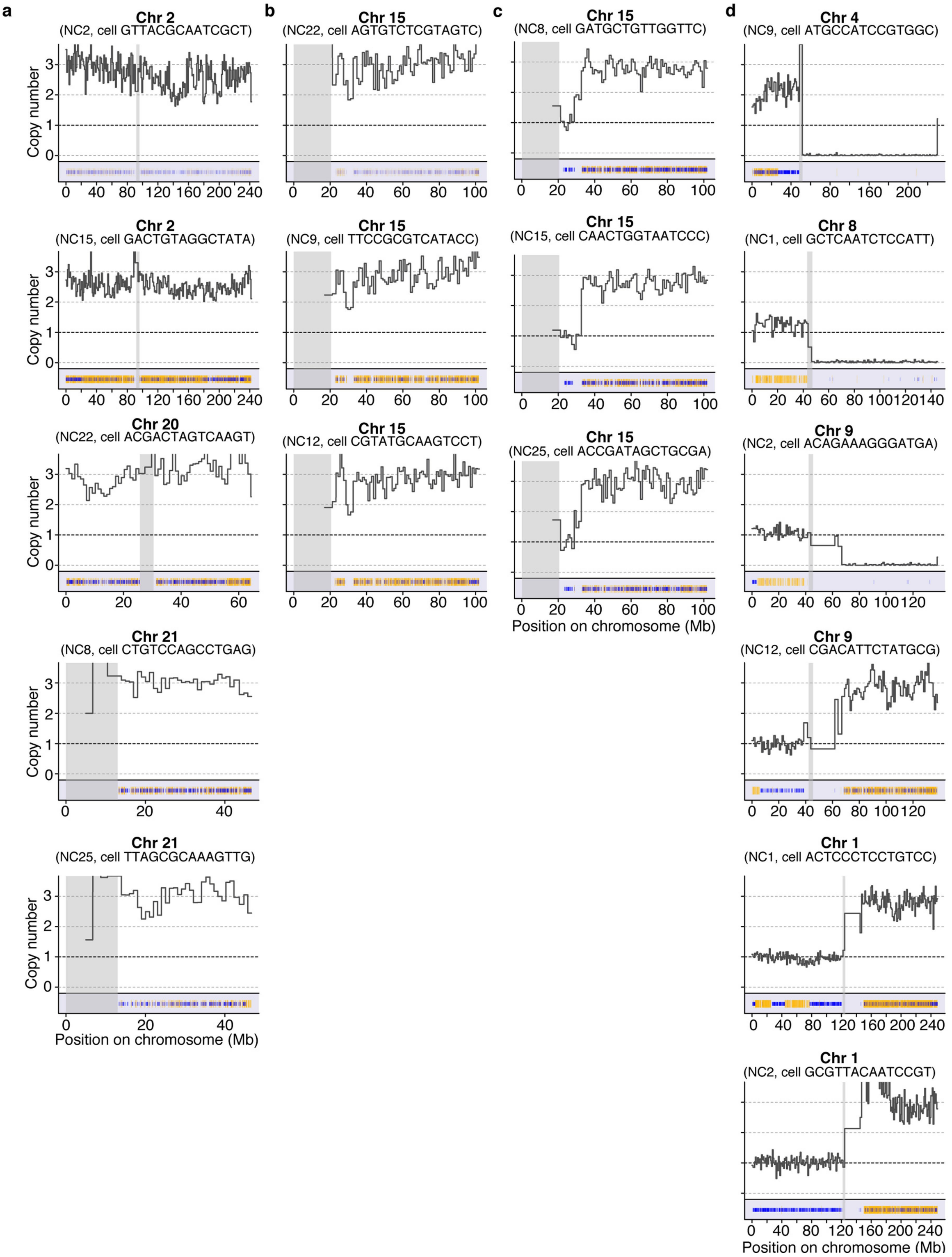
Extended Data Fig. 7 | Crossover placement in end zones, and crossover separation, varies in ways that correlate with crossover rate, among sperm donors and among individual gametes. Analyses are shown by donor (**a**–**h**; $n=20$ sperm donors) or by individual gamete (**i**, **j**, $n=31{,}228$ gametes). In **a**–**h**, the left panels show the phenotype distributions for individual donors, and the right panels show the relationship to the donors' crossover rates. To control for the effect of the number of crossovers, the analyses in **c**, **d** and **g**–**j** use 'two-crossover chromosomes'—chromosomes on which exactly two crossovers occurred. For scatter plots (**a**–**h**, right), all x-axes show the mean crossover rate and all error bars are 95% confidence intervals (y-axes are described per panel). **a**, **b**, Left, both the proportion of crossovers that falls in the most distal chromosome crossover zones (**a**) and crossover separation (**b**; a readout of crossover interference, the distance between consecutive crossovers in Mb) vary among 20 sperm donors (proportion of crossovers in end per-cell distributions among-donor Kruskal–Wallis chi-squared = 2,334, df = 19, $P<10^{-300}$; all distances between consecutive crossovers among-donor Kruskal–Wallis chi-squared = 3,309, df = 19, $P<10^{-300}$). The right panels show both properties (y-axes, total proportion of crossovers in distal zones and median crossover separation, respectively) versus the donor's crossover rate (correlation results for 20 sperm donors: proportion of all crossovers across cells in distal zones Pearson's $r=-0.95$, two-sided $P=2\times10^{-10}$; Pearson's $r=-0.96$, two-sided $P=1\times10^{-11}$). **c**, Results obtained from an alternative method for calculating the proportion of crossovers in the distal regions of chromosomes. The proportion of crossovers in the distal 50% of chromosome arms varies across donors (left, among-donor Kruskal–Wallis chi-squared = 2,209, df = 19, $P<10^{-300}$) and negatively correlates with recombination rate (right, Pearson's $r=-0.92$, two-sided $P=2\times10^{-8}$; the y-axis shows the actual proportion of crossovers in the distal 50%). **d**, As in **c**, but with the proportion of crossovers from two-crossover chromosomes occurring in the distal 50% of chromosome arms. Left, among-donor Kruskal–Wallis chi-squared = 1,058, df = 19, $P=2\times10^{-212}$; right, correlation with recombination rate Pearson's $r=-0.93$, two-sided $P=4\times10^{-9}$. **e**, As in **b** but for consecutive crossovers on the q arm of the chromosome. Left, among-donor Kruskal–Wallis chi-squared = 346, df = 19, $P=7\times10^{-62}$; right, correlation with recombination rate Pearson's $r=-0.90$, two-sided $P=5\times10^{-8}$. **f**, As in **b** but for consecutive crossovers on opposite chromosome arms (that is, crossovers that span the centromere). Left, among-donor Kruskal–Wallis chi-squared = 1,554, df = 19, $P=1<10^{-300}$; right, correlation with recombination rate Pearson's $r=-0.96$, two-sided $P=3\times10^{-11}$. **g**, As in **e** but for distances between consecutive crossovers on two-crossover chromosomes. Left, among-donor Kruskal–Wallis chi-squared = 181, df = 19, $P=2\times10^{-28}$; right, correlation with recombination rate Pearson's $r=-0.88$, two-sided $P=3\times10^{-7}$. **h**, As in **f** but for distances between consecutive crossovers on two-crossover chromosomes. Left, among-donor Kruskal–Wallis chi-squared = 930, df = 19, $P=5\times10^{-185}$; right, correlation with recombination rate Pearson's $r=-0.92$, two-sided $P=1\times10^{-8}$. **i**, **j**, Boxplots show medians and interquartile ranges with whiskers extending to 1.5 times the interquartile range from the box. Each point represents a cell. **i**, Within-donor percentiles showing the proportion of crossovers from two-crossover chromosomes that fall in distal zones, plotted against the crossover-rate decile. Groups are deciles of crossover rates normalized by converting each cell's crossover count to a percentile within-donor (all cells from all donors shown together; n cells in deciles = 3,152, 3,122, 3,276, 3,067, 3,080, 3,073, 3,135, 3,132, 3,090, 3,101, respectively (31,228 in total)). Because the initial data are proportions with small denominators, an integer effect is evident as pileups at certain values. **j**, Crossover interference from two-crossover chromosomes (showing the median consecutive crossover separation per cell). Each point represents the median of all percentile-expressed distances between crossovers from all two-crossover chromosomes in one cell (percentile taken within-chromosome); groupings and n values as in **i**.



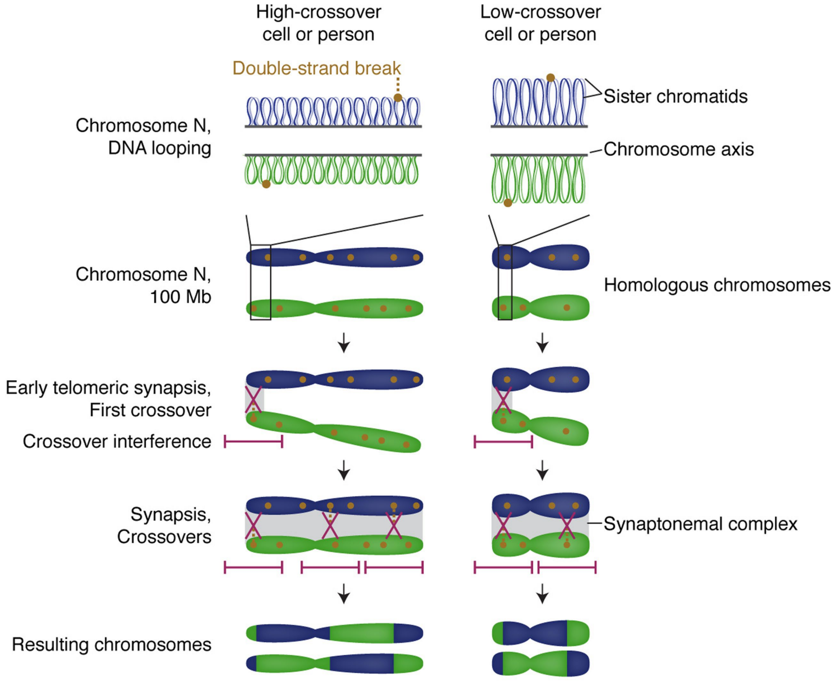
Extended Data Fig. 8 | Crossover interference in individual sperm donors and on chromosomes. a, Solid lines show density plots (scaled by donor's crossover rate) of the observed distance (separation) between consecutive crossovers as measured in the proportion of the chromosome separating them (left) and in genomic distance (right), with one line per donor ($n = 20$). Dashed lines show the distance between consecutive crossovers when crossover locations are permuted randomly across cells to remove the effect of crossover interference. **b**, The median of observed distances between consecutive crossovers for one donor (NC18, who had the tenth lowest recombination rate of 20 donors; blue dashed line) is shown along with a histogram of the medians of $n = 10,000$ among-cell crossover permutations (in both cases, the permutation one-sided P-value is less than 0.0001). The units are the proportion of the chromosome (left) and genomic distance (in Mb, right). **c**, Crossover separation on example chromosomes; plots and n values are as in **b**. Permutation one-sided $P < 0.0001$ for all chromosomes in all sperm donors except occasionally for chromosome 21, where especially few double crossovers occur. **d**, Median distances between donor NC18's consecutive crossovers for each autosome for all intercrossover distances (left two panels) and inter-crossover distances only from chromosomes with two crossovers (right two panels). Units are proportion of the chromosome or genomic distance. **e**, Diagram describing analysing crossover interference in individualized genetic distance (one 20-cM window is shown), using a donor's own recombination map. **f**, When parameterized using each donor's own genetic map, sperm donors' crossover interference profiles across multiple genetic distance windows (as shown in **e**) do not differ ($n = 20$ sperm donors; Kruskal–Wallis chi-squared = 0.22; df = 19; $P = 1$, using 20 estimates (cM distances) for each of 20 donors). Error bars show binomial 95% confidence intervals on the proportion of cells with a second crossover in the window given. This suggests that interindividual variation in crossover interference, although substantial when measured in base pairs, is negligible when measured in donor-specific genetic distance, pointing to a shared influence upon crossover interference and crossover rate.



Extended Data Fig. 9 | Relationships of aneuploidy frequency to chromosome size and recombination. a. The across-donor per-cell frequency of chromosome losses (left) and gains (centre), plotted against the length of the chromosome (from reference genome hg38; for losses across $n = 22$ chromosomes, Pearson's $r = -0.29$, two-sided $P = 0.19$; and for gains across $n = 22$ chromosomes, Pearson's $r = -0.23$, two-sided $P = 0.30$). Right, the per-chromosome rate of losses exceeding gains (number of losses minus number of gains divided by number of cells) is plotted against the length of the chromosomes (across $n = 22$ chromosomes; Pearson's $r = -0.29$, two-sided $P = 0.19$). Red labels, acrocentric chromosomes. Error bars show 95% binomial confidence intervals on the per-cell frequency (number of events/number of cells, all 31,228 cells included). **b**–**d**, Relationship between aneuploidy frequency and recombination. Only autosomal whole-chromosome aneuploidies are included. **b**, Left, total number of crossovers on meiosis I nondisjoined chromosomes (blue line; chromosomes analysed, called as transitions between the presence of one haplotype and both haplotypes on the gained chromosome) compared with $n = 10{,}000$ donor- and chromosome-matched sets (35 × 2 chromosomes per set) of properly segregated chromosomes (grey histogram; permutation). Fifty-four total crossovers on meiosis I gains versus 84.2 mean total crossovers on sets of matched chromosomes; one-sided permutation $P < 0.0001$, for the hypothesis that gained chromosomes have fewer crossovers. Right, as left but for gains occurring during meiosis II (71 meiosis-II-derived gained chromosomes of one whole copy from all individuals with fewer than five crossovers called on the gained chromosome). One-sided permutation $P = 0.98$ for meiosis II from $n = 10{,}000$ permutations, for the hypothesis that gained chromosomes have fewer crossovers; sister chromatids nondisjoined in meiosis II capture all crossovers whereas matched chromosomes do not: matched simulations and homologues nondisjoined in meiosis I capture only a random half of crossovers occurring on that chromosome in the parent spermatocyte. **c**, Crossovers per nonaneuploid megabase from each cell from each donor, split by aneuploidy status (n cells = 498, 50, 92, 30,609, left to right; 'euploid' excludes cells with any autosomal whole- or partial-chromosomal loss or gain; 'gains' includes gains of one or more than one chromosome copy; Mann–Whitney test $W = 7{,}264{,}117$, $722{,}191$, $1{,}370{,}376$; two-sided $P = 0.07$, 0.49, 0.66 for all autosomal aneuploidies, meiosis I gains and meiosis II gains, respectively, all compared against euploid). Each cell is represented by one point; boxplots show medians and interquartile ranges with whiskers extending to 1.5 times the interquartile range from the box. **d**, Per-cell crossover rates versus per-cell rates of aneuploidy (left, loss and gain; middle and right, gain only, as only chromosome gain meiotic division can be determined); $n = 20$ donors (coloured by crossover rate). P-values shown are for two-sided Pearson's correlation tests. Error bars represent 95% confidence intervals on the mean crossover rate (x-axis) and on the observed aneuploidy frequency (y-axis).



Extended Data Fig. 10 | Additional examples of noncanonical aneuploidy events detected with Sperm-seq. This figure includes those shown in Fig. 3f. Copy number, SNPs, haplotypes and centromeres are plotted as in Fig. 3a. Donor and cell identities are noted above each panel. Coordinates are in the reference genome hg38. **a**, **b**, Chromosomes 2, 20, 21 (**a**) and 15 (**b**) are sometimes present in three copies in an otherwise haploid sperm cell. **c**, A distinct, recurring triplication of much of chromosome 15, from around 33 Mb onwards but not including the proximal part of the *q* arm, also recurs in cells from three donors. **d**, Chromosome-arm-level losses (top three panels) and gains (including in more than one copy, bottom three panels, and a compound gain of the *p* arm and loss of the *q* arm, top panel).



In the example depicted:

Chromosome length			
Genomic	100 Mb	=	100 Mb
Physical	10 microns	>	5 microns
Crossover interference			
Genomic	40 Mb	<	90 Mb
Physical	4 microns	=	4 microns
Genomic proportion	0.4	<	0.9
Crossover characteristics			
Crossover number	3	>	2
Fraction of crossovers in distal regions of chromosome arms	2/3	<	1
Crossovers near centromere	Yes	>	No

Extended Data Fig. 11 | Single-cell and person-to-person variation in diverse meiotic phenotypes may be governed by variation in the physical compaction of chromosomes during meiosis. Previous work showed that the physical length of the same chromosome varies among spermatocytes at the pachytene stage of meiosis, probably by differential looping of DNA along the meiotic chromosome axis (for example, the left column shows smaller loops, resulting in more loops in total and in a greater total axis length compared with the right column, with larger loops)[15,72–75]. This physical chromosome length is correlated across chromosomes among cells from the same individual[21,76], and correlates with crossover number[15,20,21,42,73,76]. This length—measured as the length of the chromosome axis or of the synaptonemal complex (the connector of homologous chromosomes)—can vary by two or more fold among a human's spermatocytes[21]. We propose that the same process differs on average across individuals and may substantially explain interindividual variation in recombination rate. On average, individual 1 (left) would have meiotic chromosomes that are physically longer (less compacted) in an average cell than individual 2 (right); one example chromosome is shown in the figure. After the first crossover on a chromosome (probably in a distal region of a chromosome, where synapsis typically begins in male human meiosis before spreading across the whole chromosome[13–15]), crossover interference prevents nearby double-strand breaks (DSBs) from becoming crossovers; however, DSBs that are far away can become crossovers (which themselves also cause interference). More DSBs are probably created on physically longer chromosomes, and crossover interference occurs among noncrossover as well as crossover DSBs[77]. Crossover interference occurs over relatively fixed physical (micrometre) distances[43–45,76]; these distances encompass different genomic (Mb) lengths of DNA in different cells or on average in different people owing to variable compaction. Thus, crossover interference tends to lead to a different total number of crossovers as a function of the degree of compaction, resulting in the observed negative correlation (Fig. 2c, e) of crossover rate with crossover spacing (as measured in base pairs). Given that the first crossover probably occurs in a distal region of the chromosome, this model can also explain the negative correlation (Fig. 2b, d) between crossover rate and the proportion of crossovers at chromosome ends. This figure shows the total number of crossovers, crossover interference extent, and crossover locations for both sister chromatids of each homologue combined; in reality, these crossovers are distributed among the sister chromatids, making these relationships harder to detect in daughter sperm cells and requiring large numbers of observations to make relationships among these phenotypes clear.

Extended Data Table 1 | Sperm donor and single-sperm sequencing characteristics and results

Donor	Ancestry*	Cells (number excluding cell and bead doublets)	Reads per cell (median, thousands)	Genome covered per cell (median, percent)	Heterozygous SNPs in genome (millions)	Unique heterozygous SNP alleles observed per cell (median, thousands)	Crossovers observed (total, thousands)	Crossovers per cell (mean)	Resolution of crossovers (kb, median)	Autosomal aneuploidy events (percent of cells)†	Sex chromosome aneuploidy events (percent of cells)†
Overall	--	*31,228‡*	*211§*	*1.0‖*	--	*24.6§*	*813‡*	*26.11‖*	*240¶*	*1.6§*	*0.9§*
NC1	Eur.	982	284	1.4	1.95	31.6	26	26.31	189	1.5	0.6
NC2	Eur.	1,680	163	0.8	1.98	18.2	37	22.19	307	2.0	0.7
NC3	Eur.	1,289	190	0.9	1.94	21.5	36	28.13	260	1.7	0.7
NC4	Eur.	1,482	243	1.1	1.98	26.8	40	26.98	243	1.1	0.5
NC6	Afr. Am.	1,370	154	0.8	2.53	23.8	38	27.57	253	0.7	0.3
NC8	As.	1,663	304	1.5	1.81	30.9	45	26.98	229	3.1	0.5
NC9	As.	1,894	245	1.2	1.79	25.6	53	27.98	231	0.8	1.5
NC10	As.	1,154	224	1.1	1.82	23.3	29	24.99	257	1.5	0.3
NC11	Eur.	1,930	202	1.0	1.92	22.8	50	25.82	242	1.3	0.4
NC12	Eur.	2,145	179	0.9	1.91	20.6	51	23.76	270	1.2	1.7
NC13	Eur.	1,514	259	1.2	1.92	28.3	41	27.19	202	0.9	1.0
NC14	Eur.	1,336	296	1.4	1.92	32.4	36	26.65	175	2.5	1.2
NC15	Eur.	1,702	211	1.0	1.93	23.2	42	24.80	268	1.0	0.9
NC16	Eur.	1,785	241	1.2	1.92	26.9	42	23.78	227	2.2	1.3
NC17	Eur.	1,504	220	1.0	1.94	23.8	39	25.92	250	2.1	0.7
NC18	Eur.	1,589	170	0.8	1.93	18.4	42	26.48	317	1.2	0.6
NC22	Afr. Am.	1,693	195	0.9	2.53	29.7	44	25.96	205	1.4	0.7
NC25	Afr. Am.	2,274	175	0.8	2.47	25.8	62	27.31	211	2.8	1.8
NC26	Afr. Am., As.	974	120	0.6	2.55	18.0	26	26.67	355	1.3	0.4
NC27	As. (?)	1,268	267	1.3	1.96	29.2	34	26.80	199	1.7	0.6

*Ancestry as provided by the sperm bank. Afr. Am., of African American ancestry; Eur., of European ancestry; As., of Asian ancestry; (?), conflicting ancestry information given.

†These numbers are the total number of aneuploidy events divided by the total number of cells multiplied by 100; cells can have more than one event.

‡Sum across all cells from all sperm donors.

§Median or mean across all individual cells from all sperm donors (31,228 measurements summarized).

‖Median or mean of aggregate metrics across samples (20 measurements summarized).

¶Median across all crossovers (813,122 measurements summarized).

nature research

Corresponding author(s): Steven A. McCarroll

Last updated by author(s): Mar 12, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data was collected via an Illumina NovaSeq, which generated sequencing BCL files. All subsequent data processing is outlined below.
Data analysis	Picard Tools v2.2 was used for sequence data processing (http://broadinstitute.github.io/picard/). BWA-MEM v0.7.7-r441 was used for alignment (http://bio-bwa.sourceforge.net/). GATK v3.7 was used to call variants (https://gatk.broadinstitute.org/). Custom code in Drop-seq Tools v2.2 was used to format sequencing data to include single-cell barcode information and to call heterozygous data in sperm genomes (https://github.com/broadinstitute/Drop-seq/releases). Genome STRiP v2.0 was used to ascertain read depth for sperm cells (http://software.broadinstitute.org/software/genomestrip/). HapCUT v1 was used for phasing (https://github.com/vibansal/hapcut). Custom code was written in R (www.r-project.org) for this study for data processing and to call and analyze recombination and aneuploidy events; it is available via Zenodo at http://dx.doi.org/10.5281/zenodo.2581595

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Crossover and aneuploidy data (individual events and counts per donor and/or cell), including source data underlying Figs. 2, 3b-e and Extended Data Figs. 5-9, are available via Zenodo, http://dx.doi.org/10.5281/zenodo.2581570. Raw sequence data are available in the SRA via dbGaP for general research use upon application and approval (study accession number phs001887.v1.p1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We aimed to sequence ~1,000 sperm cells per individual donor to be able to detect rare events (aneuploidy occurs at a frequency of less than a few percent), and the actual number of cells sequenced per donor was a result of experimental fluctuations during library preparation. With this amount of sequencing per donor as a baseline, we then chose to sequence cells from 20 individuals to be able to observe inter-individual differences and to be able to detect trends with moderate power (for example, an analysis of 20 individuals is 80% powered to detect a Pearson's r of 0.58 at p = 0.05).
Data exclusions	No data were excluded from the analyses. For certain analyses, subsets of the data were used to examine questions about those subsets.
Replication	As this was a completely new data set, we did not seek to replicate our findings with a data set of the same type. We did compare our findings of crossover location and frequency to published estimates generated with other methods; our observations were consistent with earlier estimates (detailed in manuscript).
Randomization	N/A - we did not have experimental groups.
Blinding	N/A - we did not have experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The 20 sperm sample biospecimens analyzed in this study came from from 20 anonymous, karyotypically normal sperm donors at the New England Cryogenic Center (NECC). Per sperm bank policy, donors are 18-38 years old at the time of donation; precise age of donors is not released to researchers. The donors are known only to NECC and not to the researchers. Donor identifiers used in the paper were created specifically for this study and are not linked to any external identifiers.
Recruitment	The biospecimens used were discarded samples from New England Cryogenic Center; the samples had been collected during sperm donation. Donors consented, at the time of donation, for biospecimens to also be used for research purposes. NECC routinely provides such samples to researchers for research that has been IRB-approved at the researchers' home institution. The sperm donors might be subject to self-selection biases such as need for extra income and free time availability for donation, though any specific biases are unknown. If recombination or aneuploidy were correlated with any sperm donor recruitment or self-selection biases, the results of this study would reflect these underlying correlates. For example, the results are only reflective of donors aged 18-38 who had enough free time to act as sperm donors; if individuals of older ages or less free time had different meiotic phenotypes, they would not be captured in this study.
Ethics oversight	The Harvard Faculty of Medicine Office of Human Research Administration reviewed the research protocols (protocols M23743-101 and IRB16-0834) and determined that this research was "Not human subject research", based on the use of discarded biospecimens and the fact that researchers did not have interactions with the biospecimen donors. Harvard has also reviewed and approved our deposition of the data into an NIH respository; dbGaP will assure that access is provided only to researchers with legitimate research uses who agree not to try to re-identify donors based on data in the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Pervasive lesion segregation shapes cancer genome evolution

https://doi.org/10.1038/s41586-020-2435-1

Received: 6 December 2019

Accepted: 7 May 2020

Sarah J. Aitken[1,2,3], Craig J. Anderson[4,13], Frances Connor[1,13], Oriol Pich[5], Vasavi Sundaram[1,6], Christine Feig[1], Tim F. Rayner[1], Margus Lukk[1], Stuart Aitken[4], Juliet Luft[4], Elissavet Kentepozidou[6], Claudia Arnedo-Pac[5], Sjoerd V. Beentjes[7], Susan E. Davies[3], Ruben M. Drews[1], Ailith Ewing[4], Vera B. Kaiser[4], Ava Khamseh[4,8], Erika López-Arribillaga[5], Aisling M. Redmond[1], Javier Santoyo-Lopez[9], Inés Sentís[5], Lana Talmane[4], Andrew D. Yates[6], Liver Cancer Evolution Consortium*, Colin A. Semple[4], Núria López-Bigas[5,10,11], Paul Flicek[1,6], Duncan T. Odom[1,12] ✉ & Martin S. Taylor[4] ✉

Cancers arise through the acquisition of oncogenic mutations and grow by clonal expansion[1,2]. Here we reveal that most mutagenic DNA lesions are not resolved into a mutated DNA base pair within a single cell cycle. Instead, DNA lesions segregate, unrepaired, into daughter cells for multiple cell generations, resulting in the chromosome-scale phasing of subsequent mutations. We characterize this process in mutagen-induced mouse liver tumours and show that DNA replication across persisting lesions can produce multiple alternative alleles in successive cell divisions, thereby generating both multiallelic and combinatorial genetic diversity. The phasing of lesions enables accurate measurement of strand-biased repair processes, quantification of oncogenic selection and fine mapping of sister-chromatid-exchange events. Finally, we demonstrate that lesion segregation is a unifying property of exogenous mutagens, including UV light and chemotherapy agents in human cells and tumours, which has profound implications for the evolution and adaptation of cancer genomes.

Analysis of cancer genomes has led to the identification of many driver mutations and mutation signatures[1,3] that illustrate how environmental mutagens cause genetic damage and increase cancer risk[4,5]. The numerous patterns of mutations identified in cancer genomes reflects the temporal and spatial heterogeneity of exogenous and endogenous exposures, mutational processes and germline variation among patients. A study of diverse human cancers identified 49 distinct single-base-substitution signatures, with almost all tumours showing evidence of at least three such signatures[3].

This intrinsic heterogeneity leads to overlapping mutation signatures that make it difficult to accurately disentangle the biases of DNA damage and repair, or to interpret the dynamics of clonal evolution. We reasoned that a more controlled and genetically uniform cancer model system would overcome some of these limitations. By effectively re-running cancer evolution hundreds of times, we aimed to explore oncogenesis and mutation patterns at high resolution and with good statistical power.

We chemically induced liver tumours in postnatal day 15 (P15) male C3H/HeOuJ inbred mice (hereafter referred to as C3H mice) (Fig. 1a; $n = 104$) using a single dose of diethylnitrosamine (DEN)[6]. For comparison and validation, we replicated the study in the divergent mouse strain CAST/EiJ[7] (hereafter referred to as CAST mice) (Extended Data Fig. 1; $n = 54$).

Whole-genome sequencing (WGS) of 371 independently-evolved tumours from 104 C3H mice (Supplementary Table 1) revealed that each genome had about 60,000 (approximately 13 per Mb) somatic point mutations (Extended Data Fig. 1a), a similar level to that found in human cancers caused by exogenous mutagens such as tobacco[8] and UV exposure[9]. Insertion–deletion mutations and larger segmental changes were rare (Extended Data Fig. 1a–f). Point mutations were predominantly (76%) T→N or their complement A→N changes (where N represents any other nucleotide; Fig. 1b, Extended Data Fig. 1g–j), consistent with the long-lived thymine adduct O^4-ethyl-deoxythymidine being the principal mutagenic lesion[10]. Known driver mutations were in the EGFR–RAS–RAF pathway[6,11,12] (Fig. 1c) and usually mutually exclusive. Similar results were replicated in CAST mice (Extended Data Fig. 1j).

Chromosome-scale segregation of lesions

In each tumour, we observed multimegabase genomic segments with pronounced Watson-versus-Crick-strand asymmetry of mutations, frequently encompassing entire chromosomes (Fig. 2). We define

---

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [2]Department of Pathology, University of Cambridge, Cambridge, UK. [3]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [4]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [7]School of Mathematics and Maxwell Institute, University of Edinburgh, Edinburgh, UK. [8]Higgs Centre for Theoretical Physics, University of Edinburgh, Edinburgh, UK. [9]Edinburgh Genomics (Clinical), The University of Edinburgh, Edinburgh, UK. [10]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [12]German Cancer Research Center (DKFZ), Division of Regulatory Genomics and Cancer Evolution, Heidelberg, Germany. [13]These authors contributed equally: Craig J. Anderson, Frances Connor. *A list of members and their affiliations appears at the end of the paper. ✉e-mail: Duncan.Odom@cruk.cam.ac.uk; martin.taylor@igmm.ed.ac.uk

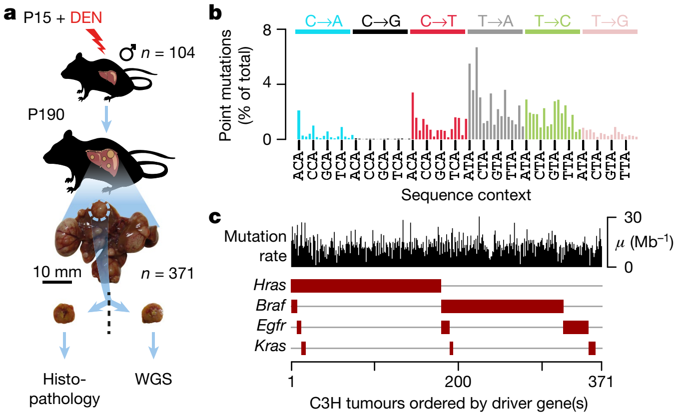


Fig. 1 | DEN-initiated tumours have a high burden of T→N and A→N mutations and driver changes in the EGFR–RAS–RAF pathway. a, P15 male C3H mice received a single dose of DEN; 371 tumours were isolated 25 weeks later (P190) and analysed by histopathology and WGS. **b**, Aggregated mutations showing the distribution of nucleotide substitutions; every fourth trinucleotide context is displayed (x-axis). **c**, Each tumour is shown as a column with its mutation rate (μ) per million base pairs (Mb) (black) and driver mutations (brown boxes).

Watson-strand bias as an excess of T→N over A→N mutations when called on the forward strand of the reference genome, and Crick-strand bias as the converse of this. With a median span of 55 Mb (Fig. 2a–d), these asymmetrically mutated segments are orders of magnitude longer than asymmetries generated by transcription-coupled nucleotide-excision repair (TCR)[13], APOBEC mutagenesis[14,15] or replication biases[13,16]. Total mutation load and DNA copy number remain uniform across the genome (Fig. 2e, f).

Pervasive, strand-asymmetric mutagenesis can be explained by DEN-induced lesions remaining unrepaired before genome replication. The first round of replication after DEN treatment results in two sister chromatids with independent lesions on each parent strand, and daughter strands containing misincorporation errors complementary to the lesions (Fig. 2i). The sister chromatids segregate into separate daughter cells during mitosis, and lesion–mutation duplexes are resolved into a mutated DNA base pair by later replication cycles. Asymmetric regions show a 23-fold excess (median) of their preferred mutation over its reverse complement, thus more than 95% of lesions that generate a mutation segregate for at least one mitotic division. We subsequently refer to this phenomenon as 'lesion segregation'.

The haploid X chromosome always contains segments with a strong strand bias (Fig. 2g). On autosomal chromosomes, when both allelic copies have the same bias, the genome shows that bias (for example, Watson bias on chromosome 15 in Fig. 2a–d); when one copy has Watson bias and the other has Crick bias, the chromosome appears unbiased (for example, chromosome 19 in Fig. 2a–d). A model based on random retention of Watson- or Crick-biased chromosomes accurately predicts that (1) around 50% of the autosomal genome and (2) 100% of the haploid X chromosome show mutational asymmetry (Fig. 2g, Extended Data Fig. 2). A few tumours (3.5%) have absent or muted asymmetry; cellularity estimates indicate that they are polyclonal or polyploid (Supplementary Table 1).

Resolving sister-chromatid exchange

The lesion segregation model predicts that mutational asymmetries should span whole chromosomes. However, we observe symmetry switches between multimegabase segments of Watson and Crick bias within chromosomes (Fig. 2a–d, g). These probably represent sister-chromatid exchanges (SCEs) from homologous-recombination-mediated DNA repair[17] (Extended Data Fig. 4a). SCEs are typically invisible to sequencing technologies because homologous recombination between sister chromatids is thought to be error-free[18].

SCE frequency per tumour positively correlates with point mutation rate (Extended Data Fig. 3a, b). With about 27 SCEs (median) in each tumour genome ($n = 371$), we had sufficient statistical power to detect recurrent exchange sites and biases in genomic context (Extended Data Fig. 3c, d). After removing three reference-genome misassemblies (Fig. 2g, Extended Data Fig. 3e, f), we found that SCEs occur with modest enrichment in transcriptionally inactive, late-replicating regions (Extended Data Fig. 4b). The fine mapping (approximately 20-kb resolution) of SCEs enabled us to test the fidelity of homologous recombination. The mutation rate appears locally elevated at SCEs, but the mutational spectrum matches the rest of the genome (Extended Data Fig. 4c–f). A model of Holliday-intermediate branch migration could explain these observations (Extended Data Fig. 4g).

Lesion segregation reveals selection

Cumulatively the tumours have equal Watson and Crick lesion-strand retention across most of the genome (Fig. 2h). However, we observe striking deviations at loci spanning known driver genes (Fig. 2h). The T→A mutation at codon 584 of the *Braf* driver gene[6] is observed in 153 out of 371 tumours in C3H mice, and we would expect the surrounding chromosomal segment to retain T lesions on the same strand. This is the case in 94% (144 out of 153) of tumours (Fisher's exact test, $P = 3.6 \times 10^{-19}$). By contrast, tumours lacking the *Braf* mutation do not show a retention bias (47% Crick, 53% Watson; $P = 0.88$, not rejecting the 50:50 null expectation). We applied this test for oncogenic selection at sites with sufficient recurrent mutations to have statistical power, which confirmed that there was significant oncogenic selection in *Hras*, *Braf* and *Egfr* (Fig. 1c, Extended Data Table 1).

DNA repair with lesion-strand resolution

Resolving DNA lesions to specific strands within a single cell cycle presents a unique opportunity to investigate strand-specific DNA damage and repair in vivo. For example, TCR (Fig. 3a) specifically removes DNA lesions from the RNA template strand[19,20].

We generated transcriptomes from the tissue of origin at the developmental time of DEN mutagenesis. Mutation rates were calculated for each gene in each tumour, stratified by both expression level and the strand containing lesions (Fig. 3b). As expected, TCR was highly specific to the template strand and correlated closely with gene expression. The mutation rate in non-expressed genes had no observable transcription-strand bias. By contrast, mutations in highly expressed genes were reduced by 79.8 ± 1.0% (mean ± s.d.) if the tumour had template-strand lesions.

To evaluate the specificity of TCR, we compared mutation rates for each trinucleotide context between template and non-template strands, stratified by expression level (Fig. 3c). For highly expressed genes, thymines have an 82 ± 6.8% (mean ± s.d. across sequence contexts) lower mutation rate on the template strand; the non-template mutation rate is indifferent to expression (Fig. 3c, dark blue lines are close to vertical), as expected[19]. Mutations from C and G show highly efficient TCR on the template strand; 70 ± 7.8% and 34 ± 21%, respectively. In contrast to T mutations, they also show an expression-dependent reduction in mutation rate on the non-template strand, suggesting that non-TCR repair processes are active. Rare mutations from adenine on the lesion-containing strand increase with transcription, possibly owing to activity of error-prone *trans*-lesion DNA polymerase Pol-η[21].

The ability to resolve the lesion strand unmasks the contribution of bidirectional transcription from active promoters[22] in shaping mutation patterns (Fig. 3d–f, Extended Data Fig. 5). Genic transcription is associated with a sharp, sustained reduction in mutation rate from template-strand lesions. A local increase in the mutation rate over

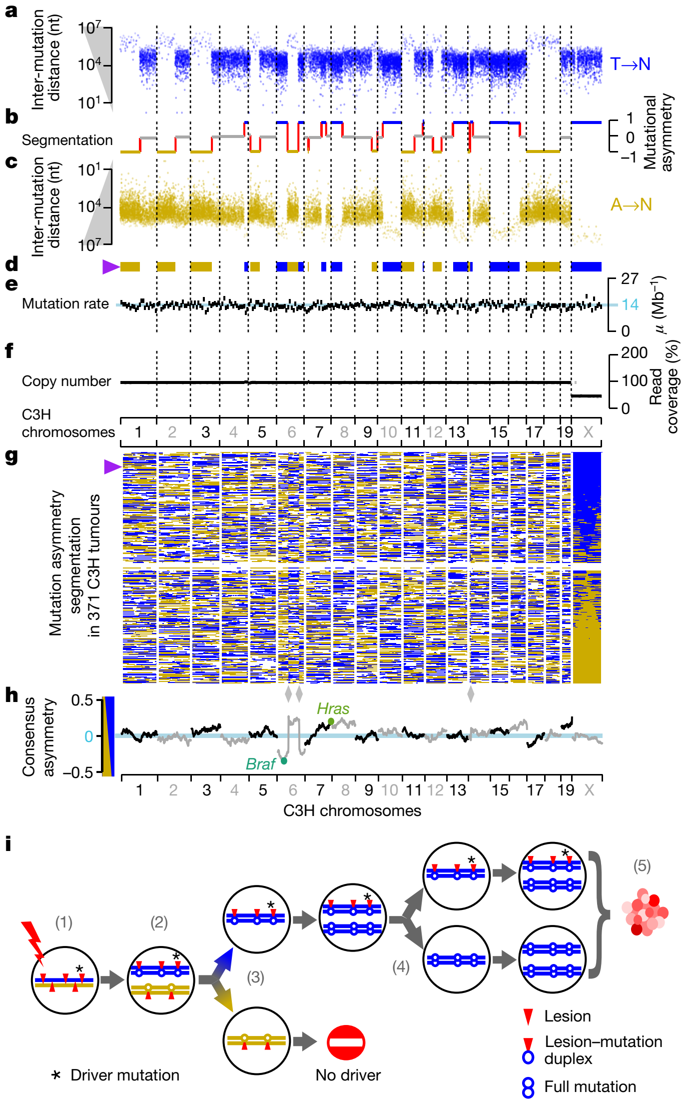


Fig. 2 | Chromosome-scale and strand asymmetric segregation of DNA lesions. **a**–**f**, An example DEN-induced C3H tumour (identifier: 94315_N8). **a**–**c**, Mutational asymmetry. Individual T→N mutations shown as blue (**a**; Watson strand) and gold (**c**; Crick strand) points; the *y*-axis shows distance to the nearest same-strand T→N mutation; nt, nucleotide. **b**, Segmentation of mutation strand asymmetry patterns. The *y*-axis shows degree of asymmetry (grey indicates no bias). Red indicates symmetry switches. **d**, Asymmetric segments shown as ribbon plot. **e**, Mutation rate in 10-Mb windows; blue line shows genome-wide average. **f**, DNA copy number in 10-Mb windows (grey) and for each asymmetric segment (black). **g**, Ribbon plots (as in **d**) for 371 C3H tumours ranked by X chromosome asymmetry. Purple triangle indicates the example tumour depicted in **a**–**f**. Grey diamonds mark reference genome misassemblies. **h**, Driver genes distort the balance of Watson and Crick asymmetries (see Methods). **i**, Mechanistic model of lesion segregation. (1) A mutagen generates lesions (red triangles) on both DNA strands. (2) If not removed, lesions will segregate into sister chromatids: one carrying only Watson-strand lesions (blue) and the second carrying only Crick-strand lesions (gold). Postmitotic daughter cells will have independent lesions and resulting replication errors (3), resolved into full mutations in later replication (4). (5) Only lineages containing driver changes (* in (1)) will expand into substantial populations.

approximately 200 nucleotides upstream of the transcription start site (Fig. 3d) is revealed to result from genic and upstream bidirectional transcription emerging from opposite edges of the core promoter[23] leading to a local depletion of TCR activity within the promoter (Fig. 3e, f).

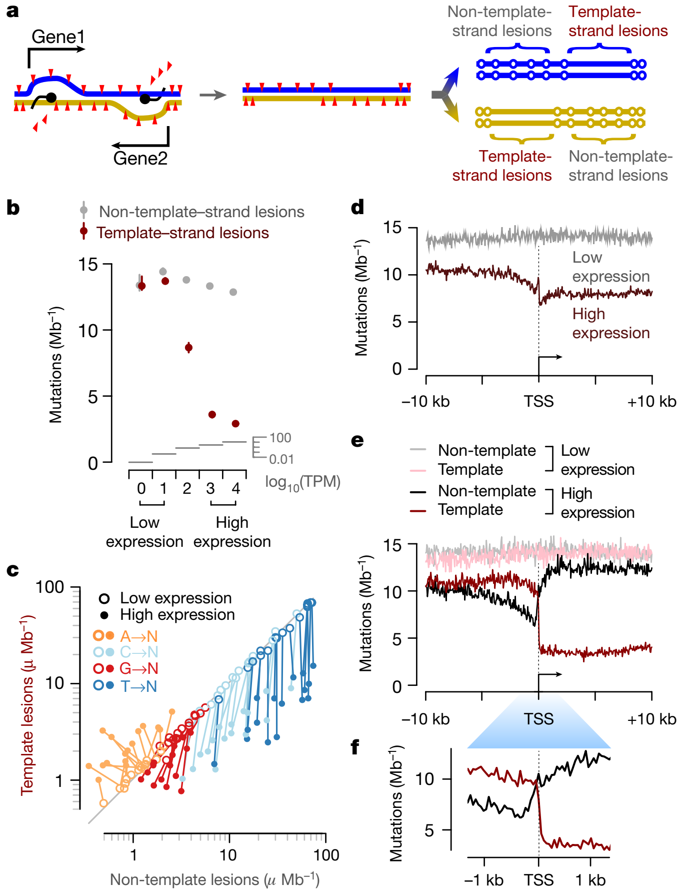


Fig. 3 | Identification of the lesion-containing DNA strand enables TCR to be quantified with strand specificity. **a**, TCR of DNA lesions is expected to reduce the mutation rate only when lesions are on the template strand of an expressed gene. Black spot shows RNA polymerase and black line the nascent RNA transcript. **b**, TCR of template-strand lesions is dependent on transcription level (P15 liver, median transcripts per million (TPM)). Estimates of mutation rate (circles) are the aggregate rates for expression level of binned genes across C3H tumours ($n = 371$). Expression level bin 0 contains $n = 2{,}835$ genes, all subsequent bins contain $n = 4{,}351 \pm 1$ genes (inclusion criteria, see Methods); empirical confidence intervals (99%) were calculated through bootstrap sampling ($n = 100$ replicates) of genes within each bin. **c**, Comparison of template versus non-template mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **d**, Sequence-composition-normalized profiles of mutation rate around transcription start sites (TSS). **e**, Stratifying by lesion strand reveals how bidirectional transcription initiation shapes the observed mutation patterns. **f**, Higher resolution of the TSS region from **e**.

An engine for genetic diversity

A segregating lesion may act as template for multiple rounds of replication in successive cell cycles (Fig. 2i). Each replication could incorporate different incorrectly or correctly paired nucleotides opposite a persistent lesion, resulting in multiple alleles at the same position. Consistent with this notion, multiallelic mutations have been reported in human cancers[24] and a cell-lineage-tracking system[25].

We evaluated multiallelic variation by identifying sites with multiple high-confidence—but conflicting—mutation calls. On average, 8% of mutated sites in DEN-induced tumours have multiallelic variants ($n = 1.8 \times 10^6$ sites in C3H tumours); per tumour, this value ranges from less than 1% to 26% (Fig. 4a). As a control, only 0.098% (95% confidence interval: 0.043–0.25%) of sites permuted between tumours show evidence of non-reference nucleotides. We further validated WGS multiallelic-variant calls using independently performed exome sequencing[6] (Fig. 4b).

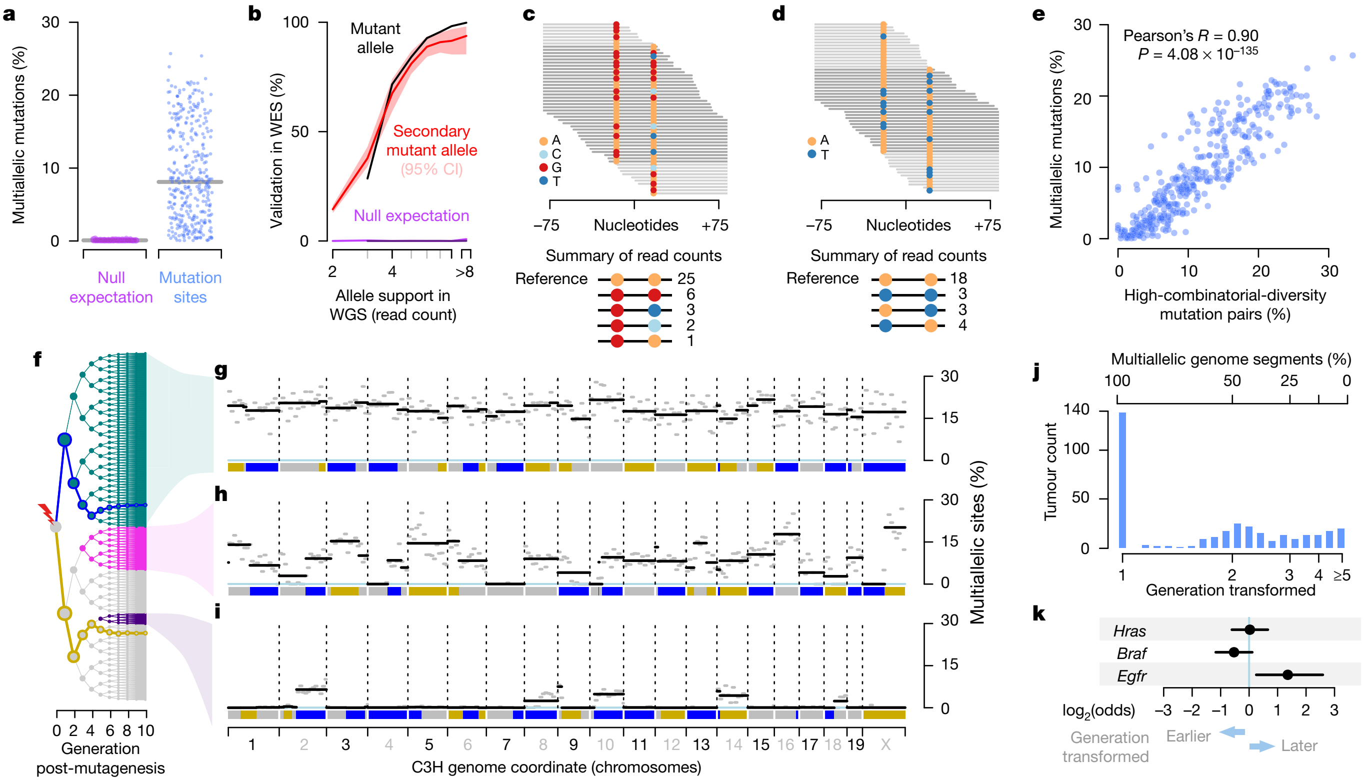


Fig. 4 | Lesion segregation generates multiallelic and combinatorial genetic diversity. a, Mutation sites per tumour with robust support for multiallelic variation; the grey line shows the median; null expectation is from permutation between tumours. **b**, Validation rate for mutations from WGS in independent whole-exome sequencing (WES); $n = 15$ tumours, collectively with $n = 20{,}683$ WGS mutations meeting inclusion criteria (Methods). Curves show validation rate stratified by WGS read support. Empirical 95% confidence interval (CI) from 100 bootstrap samplings of the aggregated WGS mutations. The null expectation permuted tumour identity between WGS and WES. **c**, Sequence reads spanning proximal mutated sites. **d**, As **c**, but showing combinatorial diversity between a pair of biallelic sites. **e**, Correlation between per-tumour multiallelic rate and combinatorially diverse mutation pairs (as in **c**, **d**), with one point per tumour. **f**, Tree of all possible progeny of a DEN-mutagenized cell for ten generations. Blue and gold lines trace simulated segregation of lesion-containing strands from a single haploid chromosome. Coloured nodes show hypothetical transformed daughter lineages with their multiallelic patterns (right). **g**–**i**, Mutation asymmetry summary ribbons for example C3H tumours that show high (**g**), variable (**h**) or low (**i**) rates of genetic diversity. The percentage of mutation sites with robust support for multiallelic variation calculated in 10-Mb windows (grey) and for each asymmetric segment (black). **j**, Histogram of the estimated cell generation post-DEN exposure from which tumours developed based on the proportion of multiallelic segments. **k**, Enrichment of specific driver gene mutations in earlier (generation 1) and later (generation >1) transforming tumours. $\log_2$ odds ratios (circles) from Fisher's exact test with 95% confidence intervals (whiskers) calculated from the hypergeometric distribution. All $n = 371$ tumours were included in the analysis for each gene.

The generation of multiallelic variation produces combinatorial genetic diversity that would not be expected under purely clonal expansion. This can be directly visualized in pairs of mutations spanned by individual sequencing reads (Fig. 4c, d). The observed combinations of biallelic sites require replication over lesions without the generation of mutations in some cell divisions (Fig. 4d). This directly demonstrates that non-mutagenic synthesis over DNA lesions occurs, and allele frequency analysis indicates it is common (Extended Data Fig. 6). The per-tumour rates of combinatorial diversity and multiallelic sites correlate closely and highlight the wide variation between tumours (Fig. 4e).

The explanation for such intertumour variance becomes evident when plotting the distribution of multiallelic sites along each genome (Fig. 4f–i). Tumours with high rates of genetic diversity have consistently high rates of multiallelism throughout their genome (Fig. 4g). They are likely to have expanded from a first-generation daughter of the original DEN-mutagenized cell, in which all DNA is a duplex of a lesion-containing and non-lesion-containing strand. Therefore, replication using lesion-containing strands as the template in subsequent generations produces multiallelic variation uniformly across the genome. Tumours with lower total levels of genetic diversity exhibit discrete genomic segments of high and low multiallelism (Fig. 4h, i). These tumours probably developed from a cell some generations after DEN treatment. Each mitosis following DEN exposure is expected to dilute the number of lesion-containing strands in each daughter cell by approximately 50%. Only lesion-retaining fractions of the genome generate multiallelic and combinatorial genetic diversity in the daughter lineages; consistent with this, the multiallelic segments mirror the mutational asymmetry segmentation pattern.

By estimating the fraction of multiallelic chromosomal segments, we can infer the cell generation, relative to DEN exposure, from which the tumour expanded (Fig. 4j). In 67% of C3H tumours and 21% of CAST tumours, the initial burst of mutations was instantly transformative. In the remainder of tumours, the observed fractions of multiallelic segments cluster around expectations for subsequent cell generations, suggesting that transformation required a specific combination of mutated alleles, an additional mutation or an external trigger. Of note, *Egfr*-driven tumours appear to transform significantly later ($P = 0.042$ after Bonferroni correction, Fisher's exact test), suggesting that driver gene identity influences the timing of tumour inception (Fig. 4k).

Lesion segregation is ubiquitous

Lesion segregation is a feature of DEN mutagenesis in mice. This raises two critical questions. Do other DNA-damaging agents induce lesion

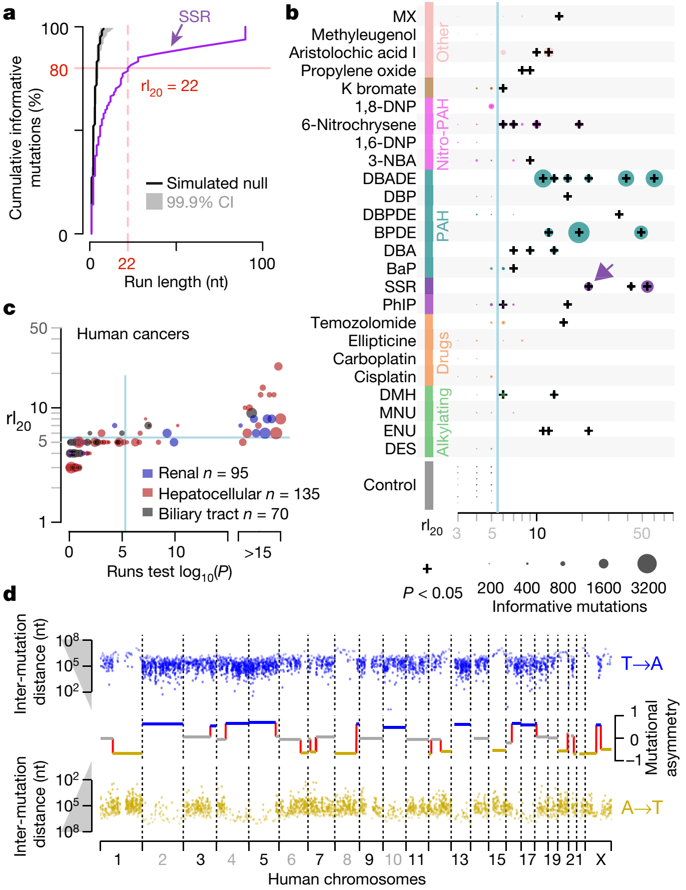


Fig. 5 | Lesion segregation is a pervasive feature of exogenous mutagens and is evident in human cancers. a, The runs-based rl_{20} metric, calculated for an example simulated solar radiation (SSR) clone (Extended Data Fig. 7a); 20% of informative mutations (C→T or their complement G→A) are in strand asymmetric runs of at least 22 consecutive mutations (for example, ≥22 C→T mutations without an intervening G→A). Simulated null based on 100,000 permutations of 1,000 mutations; black curve shows median. **b**, All robust mutagens in human iPSCs[5], mutagen classes indicated by coloured boxes; PAH indicates polycyclic aromatic hydrocarbons. Individual compound abbreviations expanded in Supplementary Table 2. The rl_{20} metric (x-axis) is plotted for each clone ($n = 325$), including multiple replicates per exposure. Data point size quantifies informative mutations; $^{+}P < 0.05$ (two-sided, Bonferroni-corrected). **c**, The rl_{20} metric and runs tests for human cancers[27]; $n = 18{,}850$ cancers screened, three cohorts plotted. Blue lines show Bonferroni adjusted $P = 0.05$ threshold for the runs test (two-sided) and an empirical threshold for rl_{20} (Methods). x-axis P-values $<1 \times 10^{-15}$ are rank-ordered. **d**, Mutational asymmetry (plotted as in Fig. 2a–c) in a human hepatocellular carcinoma (donor DO231953) with a dominant mutation signature for aristolochic acid exposure.

segregation? Does lesion segregation occur in human cells and cancers? Recently, a study in which human induced pluripotent stem cells (iPSCs) were exposed to 79 environmental mutagens revealed that 41 of the mutagens produced excess nucleotide substitutions[5]. Although not previously noted in these in vitro data, many of the exposures generated chromosome-scale lesion segregation patterns (Extended Data Fig. 7) similar to those observed in the in vivo DEN model. Applying runs-based tests (Fig. 5a, b, Extended Data Fig. 8), we detect marked mutational asymmetry in every sample with more than 1,000 'informative' mutations (Fig. 5b, Extended Data Fig. 8b; see Methods), including clinically relevant insults such as sunlight (simulated solar radiation), tobacco smoke (benzo[*a*]pyrene diol-epoxide (BPDE)) and chemotherapeutics (temozolomide). By contrast, mutations induced by perturbation of replication and repair pathways[26] independent of DNA lesions showed no detectable asymmetry, as expected (Extended Data Fig. 8c). We conclude that the chromosome-scale segregation of lesions and the resulting strand asymmetry of mutation patterns, are general features of all tested DNA-damaging mutagens.

The pronounced mutational asymmetry observed in both DEN-induced tumours and mutagen-exposed human iPSCs[5] occurs after a single mutagenic insult. By contrast, most human cancers accumulate mutations as a result of multiple damaging events over their history. Lesion segregation predicts that such tumours will acquire new waves of segregating lesions after each exposure, thus progressively masking their asymmetry patterns. Therefore, even though UV exposure causes substantial lesion segregation in human cells (Fig. 5a, b, Extended Data Figs. 7a, 8b), it is unlikely that skin cancers would show mutational asymmetry following repeated UV exposure.

Nevertheless, analysis of human cancer genomes[27] ($n = 18{,}850$ tumours, 22 primary sites) identified multiple cancers with the characteristic mutational asymmetry of lesion segregation (Fig. 5c, d). The majority of these tumours are renal, hepatic or biliary in origin, and show a high mutation rate and strand asymmetry of T→A or their complement A→T mutations, consistent with exposure to aristolochic acid[3] (Supplementary Table 2). Although it is seen most clearly in tumours subjected to a single dose of a mutagen, lesion segregation probably shapes all genomes subjected to DNA damage, with important implications for tumour evolution and heterogeneity.

Discussion

In this study, we have shown that most mutation-causing DNA lesions are not resolved as mutations within a single cell cycle. Instead, lesions segregate unrepaired into daughter cells for multiple cellular generations, resulting in chromosome-scale strand asymmetry of subsequent mutations. This suggests that lesion removal before replication has high fidelity and rarely results in mutations. Lesion segregation was initially discovered in an in vivo mouse model of oncogenesis; we have demonstrated that it is ubiquitous for all tested mutagens, also occurs in human cells and is evident in human cancers. Similar patterns of asymmetry in bacterial mutagenesis suggest that the underlying mechanisms are highly conserved[28,29].

Our discovery of lesion segregation challenges longstanding assumptions of cancer evolution[30]. For example, the widely used infinite sites model[31] does not allow for recurrent mutation at the same site. Our findings also provide new perspectives for understanding cancer evolution using mutational asymmetry and multiallelism patterns to track events during oncogenesis and to quantify selection. Perhaps most notably, lesion segregation is a previously unrecognized mechanism for a cancer to sample the fitness effects of mutation combinations, thus evading Muller's ratchet[32] and Hill–Robertson interference, which assumes low selection efficiency owing to the inability to separate mutations of opposing fitness[33,34]. Consequently, DNA-damaging chemotherapeutics, particularly large or closely spaced doses generating persistent lesions, could inadvertently provide an opportunity for cancer to efficiently select resulting mutations. This insight may guide the development of more effective chemotherapeutic regimens.

Once identified, lesion segregation is a deeply intuitive concept. Its practical applications provide new vistas for the exploration of genome maintenance and fundamental molecular biology. The discovery of pervasive lesion segregation profoundly revises our understanding of how the architecture of DNA repair and clonal proliferation can conspire to shape the cancer genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2435-1.

1. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet*. **20**, 404–416 (2019).
3. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
5. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
6. Connor, F. et al. Mutational landscape of a chemically-induced mouse model of liver cancer. *J. Hepatol*. **69**, 840–850 (2018).
7. Maronpot, R. R. Biological basis of differential susceptibility to hepatocarcinogenesis among mouse strains. *J. Toxicol. Pathol*. **22**, 11–33 (2009).
8. Wang, C. et al. Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat. Commun*. **9**, 2054 (2018).
9. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
10. Verna, L., Whysner, J. & Williams, G. M. *N*-nitrosodiethylamine mechanistic data and risk assessment: bioactivation, DNA-adduct formation, mutagenicity, and tumor initiation. *Pharmacol. Ther*. **71**, 57–81 (1996).
11. Maronpot, R. R., Fox, T., Malarkey, D. E. & Goldsworthy, T. L. Mutations in the ras proto-oncogene: clues to etiology and molecular pathogenesis of mouse liver tumors. *Toxicology* **101**, 125–156 (1995).
12. Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J. & Schwarz, M. Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutat. Res*. **638**, 66–74 (2008).
13. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
14. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
15. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294 (2019).
16. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol*. **19**, 129 (2018).
17. Perry, P. & Evans, H. J. Cytological detection of mutagen-carcinogen exposure by sister chromatid exchange. *Nature* **258**, 121–125 (1975).
18. Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Front. Genet*. **5**, 175 (2014).
19. Strick, T. R. & Portman, J. R. Transcription-coupled repair: from cells to single molecules and back again. *J. Mol. Biol*. **431**, 4093–4102 (2019).
20. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev*. **29**, 948–960 (2015).
21. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
22. Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
23. Preker, P. et al. PROMoter uPstream transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res*. **39**, 7179–7193 (2011).
24. Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res*. **27**, 1885–1894 (2017).
25. Brody, Y. et al. Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Res*. **28**, 1901–1918 (2018).
26. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun*. **9**, 1744 (2018).
27. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
28. Parkhomchuk, D., Amstislavskiy, V., Soldatov, A. & Ogryzko, V. Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proc. Natl Acad. Sci. USA* **106**, 20830–20835 (2009).
29. Chan, K. & Gordenin, D. A. Clusters of multiple mutations: incidence and molecular mechanisms. *Annu. Rev. Genet*. **49**, 243–267 (2015).
30. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet*. **18**, 213–229 (2017).
31. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
32. Zhang, Y. et al. Genetic load and potential mutational meltdown in cancer cell populations. *Mol. Biol. Evol*. **36**, 541–552 (2019).
33. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res*. **8**, 269–294 (1966).
34. Tilk, S., Curtis, C., Petrov, D. & McFarland, C. D. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. Preprint at *bioRxiv* https://doi.org/10.1101/764340 (2019).

Liver Cancer Evolution Consortium

Sarah J. Aitken[1,2,3], Stuart Aitken[4], Craig J. Anderson[4], Claudia Arnedo-Pac[5], Frances Connor[1], Ruben M. Drews[1], Ailith Ewing[4], Christine Feig[1], Paul Flicek[1,6], Vera B. Kaiser[4], Elissavet Kentepozidou[6], Erika López-Arribillaga[5], Núria López-Bigas[5,10,11], Juliet Luft[4], Margus Lukk[1], Duncan T. Odom[1,12]✉, Oriol Pich[5], Tim F. Rayner[1], Colin A. Semple[4], Inés Sentís[5], Vasavi Sundaram[1,6], Lana Talmane[4] & Martin S. Taylor[4]✉

Methods

Statistical methods were used to predetermine sample size for the testing of oncogenic selection by biased strand retention; otherwise, no statistical methods were used to determine sample size. The investigators were blinded to allocation during histopathological assessment.

Mouse colony management

Animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical Review Body (AWERB): the maximum approved tumour burden was 10% body weight, which was not exceeded. Animals were maintained using standard husbandry: mice were group housed in Tecniplast GM500 IVC cages with a 12 h:12 h light:dark cycle and ad libitum access to water, food (LabDiet 5058), and environmental enrichments.

Chemical model of hepatocarcinogenesis

P15 male C3H and CAST mice were treated with a single intraperitoneal (IP) injection of DEN (Sigma-Aldrich N0258; 20 mg kg^{-1} body weight) diluted in 0.85% saline. Liver tumour samples were collected from DEN-treated mice 25 weeks (C3H) or 38 weeks (CAST) after treatment. All macroscopically identified tumours were isolated and processed in parallel for DNA extraction and histopathological examination. Non-tumour tissue from untreated P15 mice (ear, tail, and background liver) was sampled for control experiments.

Tissue collection and processing

Liver tumours of sufficient size (≥2 mm diameter) were bisected; one half was flash frozen in liquid nitrogen and stored at −80 °C for DNA extraction, and the other half was processed for histology. Tissue samples for histology were fixed in 10% neutral buffered formalin for 24 h, transferred to 70% ethanol, machine processed (Leica ASP300 Tissue Processor; Leica), and paraffin embedded. All formalin-fixed paraffin-embedded sections were 3 µm in thickness.

Histochemical staining

Formalin-fixed paraffin-embedded tissue sections were stained with haematoxylin and eosin (H&E) using standard laboratory techniques. Histochemical staining was performed using the automated Leica ST5020; mounting was performed on the Leica CV5030.

Imaging

Tissue sections were digitised using the Aperio XT system (Leica Biosystems) at 20× resolution; all H&E images are available in the BioStudies archive at EMBL-EBI under accession S-BSST383.

Tumour histopathology

H&E sections of liver tumours were blinded and assessed twice by a pathologist (S.J.A.); discordant results were reviewed by an independent hepatobiliary pathologist (S.E.D.). Tumours were classified according to the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) guidelines for lesions in rats and mice[35]. In addition, tumour grade, size, morphological subtype, nature of steatosis and mitotic index were assessed (Supplementary Table 1), as well as the presence of cystic change, haemorrhage, necrosis, or vascular invasion.

Sample selection for WGS

Tumours which met the following histological criteria were selected for WGS (C3H $n = 371$, CAST $n = 84$): (i) diagnosis of either dysplastic nodule (DN) or hepatocellular carcinoma (HCC), (ii) homogenous tumour morphology, (iii) tumour cell percentage >70%, and (iv) adequate tissue for DNA extraction. Neoplasms with extensive necrosis, mixed tumour types, a nodule-in-nodule appearance (indicative of an HCC arising within a DN), or contamination by normal liver tissue were excluded. Since carcinogen-induced tumours arising in the same liver are independent[6], multiple tumours were selected from each mouse to minimise the number of animals used. A subset of normal (non-tumour) samples from untreated mice were also sequenced (C3H $n = 13$, CAST $n = 7$).

Whole-genome sequencing

Genomic DNA was isolated from liver tissue and liver tumours using the AllPrep 96 DNA/RNA Kit (Qiagen, 80311) according to the manufacturer's instructions. DNA quality was assessed on a 1% agarose gel and quantified using the Quant-IT dsDNA Broad Range Kit (Thermo Fisher Scientific). Genomic DNA was sheared using a Covaris LE220 focused-ultrasonicator to a 450-bp mean insert size.

WGS libraries were generated from 1 µg of 50 ng μl^{-1} high molecular weight genomic DNA using the TruSeq PCR-free Library Prep Kit (Illumina), according to the manufacturer's instructions. Library fragment size was determined using a Caliper GX Touch with a HT DNA 1k/12K/Hi Sensitivity LabChip and HT DNA Hi Sensitivity Reagent Kit to ensure fragments of 300–800 bp (target ~450 bp).

Libraries were quantified by real-time PCR using the Kapa library quantification kit (Kapa Biosystems) on a Roche LightCycler 480. 0.75 nM libraries were pooled in 6-plex and sequenced on a HiSeq X Ten (Illumina) to produce paired-end 150-bp reads. Each pool of 6 libraries was sequenced over eight lanes (minimum of 40× coverage).

Variant calling and somatic mutation filtering

Sequencing reads were aligned to respective genome assemblies (C3H = C3H_HeJ_v1; CAST = CAST_EiJ_v1)[36] with bwa-mem (v.0.7.12)[37] using default parameters. Reads were annotated to read groups using the Picard (v.1.124)[38] tool AddOrReplaceReadGroups, and minor annotation inconsistencies corrected using the Picard CleanSam and FixMateInformation tools. Bam files were merged as necessary, and duplicate reads were annotated using the Picard tool MarkDuplicates.

Single-nucleotide variants were called using Strelka2 (v.2.8.4)[39] implementing default parameters. Initial variant annotation was performed with the GATK (v.3.8.0)[40] walker CalculateSNVMetrics (https://github.com/crukci-bioinformatics/gatk-tools). Genotype calls with a variant allele frequency <0.025 were removed. Although inbred strains were used, fixed genetic differences between the colonies and the reference genome, as well as small numbers of germline variants segregating within the colonies were identified. For each strain, fixed differences identified as homozygous changes present in 100% of genotyped samples were filtered out. Segregating variants were filtered based on the excess clustering of mutations to animals with shared mothers. To generate a null expectation taking into account the family structure of the colonies, the parent–offspring relationships were randomly permuted 1,000 times. For each count of recurrent mutation (range 5 to 371 inclusive), we determined the null distribution of expected distinct mothers. Comparing this to the observed count of distinct mothers for each recurrent ($n > 4$) mutation, those with a low probability ($P < 1\text{x}10^{-4}$, pnorm function from R (v.3.5.1)[41]) under the null were excluded from analyses.

Copy number variation between tumours within strains was called using CNVkit (v.0.9.6)[42]. Non-tumour reference coverage was provided from non-tumour control WGS data (C3H $n = 11$, CAST $n = 7$) and per tumour cellularity estimates (see below) were provided.

RNA sequencing

Total RNA was extracted from P15 liver tissue ($n = 4$ biological replicates per strain) using QIAzol Lysis Reagent (Qiagen), according to manufacturer's instructions. DNase treatment and removal were performed using the TURBO DNA-free kit (Ambion, Life Technologies), according to the manufacturer's instructions. RNA concentration was measured using a NanoDrop spectrophotometer (Thermo Fisher); RNA integrity was assessed on a Total RNA Nano Chip Bioanalyzer (Agilent).

Total RNA (1 µg) was used to generate sequencing libraries using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina), according to the manufacturer's instructions. Library fragment size was determined using a 2100 Bioanalyzer (Agilent). Libraries were quantified by quantitative PCR (Kapa Biosystems). Pooled libraries were sequenced on a HiSeq4000 to produce ≥ 40 million paired-end 150-bp reads per library.

RNA-seq data processing and analysis

Transcript abundances were quantified with Kallisto (v.0.43.1)[43] (using the flag–bias) and a transcriptome index compiled from coding and non-coding cDNA sequences defined in Ensembl v91[44]. TPM estimates were generated for each annotated transcript and summed across alternate transcripts of the same gene for gene-level analysis. The TSS for each gene was annotated with Ensembl v91 and based upon the most abundantly expressed transcript. RNA-sequencing (RNA-seq) data are available at Array Express at EMBL-EBI under accession E-MTAB-8518.

Genomic annotation data

Mouse liver proximity ligation sequencing (HiC) data were downloaded from GEO (GSE65126)[45], replicates were combined, then aligned to GRCm38[46] and processed using the Juicebox (v.7.5) and Juicer scripts[47] to obtain the HiC matrix. Eigenvectors were obtained for 500kb consecutive genomic windows over each chromosome from the HiC matrix using Juicebox and subsequently oriented (to distinguish compartment A from B) using GC content per 500-kb bin. We used progressiveCactus[48] to project the 500-kb windows into the C3H reference genome and Bedtools (v.2.28.0) to merge syntenic loci between 450 and 550 kb in size, removing the second instance where we observed overlaps.

Genic annotation was obtained from Ensembl v.91[44] for the corresponding C3H and CAST reference genome assemblies (C3H_HeJ_v1, CAST_EiJ_v1). Genomic repeat elements were annotated using RepeatMasker (v.20170127; http://www.repeatmasker.org) with the default parameters and libraries for mouse annotation.

The analysable fraction of the genome

Analysis and sequence composition calculations were confined to the main chromosome assemblies of the reference genome (chromosomes 1–19 and X). Using WGS of non-tumour liver, ear and tail samples (C3H $n = 11$, CAST $n = 7$) collected and sequenced contemporaneously with tumour samples, genome sequencing coverage was calculated for 1-kb windows using multicov in Bedtools (v.2.28.0)[49]. Windows with read coverage >2 s.d. from the autosomal mean were flagged as suspect in each tumour. Read coverage over the X chromosome was doubled in these calculations to account for the expected hemizygosity in these male mice. Any 1-kb window identified as suspect in >90% of these non-tumour samples was flagged as 'abnormal read coverage' (ARC) and masked from subsequent analysis. This masked 12.7% of the C3H and 11.5% of the CAST reference genomes yielding analysable haploid genome sizes of C3H = 2,333,783,789 nucleotides (nt) and CAST = 2,331,370,397 nt.

Mutation rate calculations

Mutation rates were calculated as 192 category vectors representing every possible single-nucleotide substitution conditioned on the identity of the upstream and downstream nucleotides. Each rate being the observed count of a mutation category divided by the count of the trinucleotide context in the analysed sequence. To report a single aggregate mutation rate, the three rates for each trinucleotide context were summed to give a 64 category vector and the weighted mean of that vector reported as the mutation rate. The vector of weights being the trinucleotide sequence frequency of a reference sequence, for example the composition of the whole genome. In the case of whole-genome analysis, the same trinucleotide counts are used in (1) the individual category rates calculation and (2) the weighted mean of the rates, cancelling out. For windowed comparisons of mutation rates, the weighted mean is calculated using the genome wide composition of trinucleotides rather than the local sequence composition, providing a compositionally adjusted mutation rate estimate. For mutation rates in TCR analysis, the same compositional adjustment was carried out but using the trinucleotide composition of the aggregate genic spans of genome (minus ARC regions) for normalization.

Mutation signatures

The 96 category 'folded' mutation counts for each of the 371 C3H tumours were deconvolved into the best fitting number (K) of component signatures using sigFit (v.2.0)[50] with 1,000 iterations and K set to integers 2 to 8 inclusive. A heuristic goodness-of-fit score based on cosine similarity favoured instances where $K = 2$. The DEN1 and DEN2 signatures reported were obtained by running sigFit with 30,000 iterations for $K = 2$. Analysis of CAST tumours gave less distinct separation of signatures so the C3H derived DEN1 and DEN2 were used for both strains. To fit signatures to each tumour we used sigFit provided with the DEN signatures and additional SPONT1 and SPONT2 signatures that were derived from equivalent WGS analysis of spontaneous (non-DEN-induced) C3H tumours.

Driver mutation identification

Candidate cancer driver genes were identified by applying OncodriveFML (v.2.2.0 using the SIFT scoring scheme)[51] and OncodriveCLUSTL (v.1.1.1)[52] to mutations identified in C3H tumours. The only genes convincingly identified as significantly enriched for functionally impactful or clustered mutations were *Braf*, *Egfr* and *Hras*. *Kras* appeared as marginally significant. These four genes were identified for C3H[6]. Protein altering mutations in those genes were annotated as driver mutations in C3H and CAST tumours.

Mutational asymmetry segmentation and scoring

For each tumour a focal subset of 'informative' mutation types were defined, T→N or A→N mutations, in the case of DEN-induced tumours. The order of focal mutations along each chromosome was represented as a binary vector (for example, 0 for T→N, 1 for A→N). Vectors corresponding to each chromosome of each tumour were processed with the cpt.mean function of the R Changepoint (v.2.2.2)[53] package run with an Akaike information criterion (AIC) penalty function, maximum number of changepoints set to 12 ($Q = 12$), and implementing the PELT algorithm for optimal changepoint detection. Following segmentation, the defined segments were scored for strand asymmetry, taking into account the sequence composition of the segment. For example in tumours with T→N or A→N informative mutations the number of Ts on the forward strand is the count of Watson sites G_W and the number of T→N mutations is μ_W which together give the Watson strand rate $R_W = \mu_W/G_W$. The forward strand count of As and mutations from A likewise give the Crick strand rate $R_C = \mu_C/G_C$. From these two rates we calculate a relative difference metric, the mutational asymmetry score $S = (R_W - R_C)/(R_W + R_C)$.

The parameter S scales from 1 all Watson (for example, DEN T→N mutations) through 0 (50:50 T→N:A→N) to −1 for all Crick (for example, DEN A→N). For the categorical assignment, $S \geq 0.3$ is Watson-strand asymmetric, $S \leq -0.3$ Crick-strand asymmetric and in the range $-0.3 < S < 0.3$ symmetric, though more stringent filtering was applied where noted. Segments containing <20 informative mutations were discarded from subsequent analyses.

To test for oncogenic selection at sites with recurrent mutations, mutational asymmetry segments overlapping the focal mutation were categorised based on their asymmetry score S, as above. The test was implemented as a Fisher's exact test with the 2 × 2 contingency table comprising the counts chromosomes (two autosomes per cell) stratified by Watson-versus-Crick asymmetry and the presence of the focal mutation in the tumour. Tumours containing another known driver gene or recurrent mutation within the focal asymmetry segment were discarded from the analysis. We estimated the minimum recurrence of a mutation necessary to reliably detect oncogenic selection through simulation. Biased segregation of chromosomes containing drivers was modelled using the observed median excess of T→N over A→N lesions (23-fold), and random segregation of non-driver containing strands (1:1 ratio). Our model predicted >33 C3H recurrences or >41 CAST recurrences would give 80% power to detect oncogenic selection if present.

Tumour cellularity estimates

We calculated tumour cellularity as a function of the non-reference read count in autosomal chromosomes $(1 - R/d) \times 2$, where R is the reference read count at a mutated site and d is the total read depth at the site. For each tumour these values were binned in percentiles and the midpoint of the most populated (modal) percentile taken as the estimated cellularity of the tumour. Given the low rate of copy number variation across the DEN induced tumours, no correction was made for copy-number distortion. Skew in the variant allele frequency (VAF) = $(1 - R/d)$) distribution was calculated using Pearson's median skewness coefficient implemented in R as $(3 \times (\text{mean} - \text{median}))/\text{s.d.}$ of the VAF distribution.

Identifying and filtering reference genome misassemblies

Since lesion segregation, mutation asymmetry patterns allow the long-range phasing of chromosome strands, they can detect discrepancies in sequence order and orientation between the sequenced genomes and the reference. We identified autosomal asymmetry segments that immediately transitioned from Watson bias ($S > 0.3$) to Crick ($S < -0.3$) or vice versa without occupying the intermediate unbiased state ($-0.3 < S < 0.3$); such discordant segments are unexpected. Allowing for ±100 kb uncertainty in the position of each exchange site we produced the discordant segment coverage metric. At sites with discordant segment coverage >1 we calculated percentage consensus for misassembly $M = \text{ds}/(\text{ds}+\text{cs})$ where ds is the number of discordant segments over the exchange site and cs the number of concordant: where either Watson or Crick mutational asymmetry extends at least 1×10^6 nucleotides on both sides of the exchange site. The approximate genomic coordinates for a C3H strain specific inversion on chromosome 6 have been previously reported[54].

SCE-site analysis

Identified SCE sites were aggregated across tumours from each strain. Exchange sites within 1×10^6 nt of known and proposed reference genome misassembly sites were excluded from analysis. The mid-point between the flanking informative mutations was taken as the reference genome position of the exchange event, and the distance between those flanking mutations as the positional uncertainty of the estimate. To generate null expectations for mutation rate measures, the coordinate of an exchange was projected into the genome of a proxy tumour and the mutation rates and patterns measured from that proxy tumour (repeated 100 times). The permutation of tumour identifiers for the selection of proxy tumours was a shuffle without replacement that preserved the total number of exchange sites measured in each tumour.

The comparison of mutation spectra between windows was calculated as the cosine distance between the 96 category trinucleotide context mutation spectra for the whole genome and that calculated for the aggregated 5-kb window. The 96 categories were equally weighted for this comparison.

Exchange site enrichment analysis used Bedtools[49] shuffle to permute the genomic positions of exchange sites into the analysable fraction of the genome (defined above). Observed rates of annotation overlap were compared to the distribution of values from 1,000 permuted exchange sites. For genic overlaps we used Ensembl v.91[44] coordinates for genic spans; gene expression status was based on the summed expression over all annotated transcripts for the gene from P15 liver from the matched mouse strain. Expression thresholds were defined as >50th centile for active and <50th centile for inactive genes.

A higher count of informative mutations provides greater power to identify shorter mutational asymmetry segments. To fairly test for correlation between nucleotide substitution rate and SCE rate we randomly down-sampled informative mutations to 10,000 per tumour genome and recomputed the mutational asymmetry segmentation patterns from the sampled data. Tumours with <10,000 informative mutations were excluded. We then correlated the total (not down sampled) nucleotide substitution load to the count of SCE events inferred from the down-sampled data.

TCR calculations

For each protein coding gene, the maximally expressed transcript isoform was identified from P15 liver in the matched strain (TPM expression), subsequently the primary transcripts. In the case of ties, transcript selection was arbitrary. Genes were partitioned into five categories based on the expression of the primary transcript: expression level 0 (<0.0001 TPM) and four quartiles of detected expression.

Using the segmental asymmetry patterns of each tumour and the annotated coordinates (Ensembl v.91) of the selected transcripts, we identified transcripts completely contained in a single Watson or Crick asymmetric segment and located at least 200 kb from the segment boundary at both ends. We also applied strict asymmetry criteria of mutational asymmetry scores $S > 0.8$ for Watson and $S < -0.8$ for Crick asymmetry segments, though analysis with the standard asymmetry thresholds and no segment boundary margin give similar results and identical conclusions. For each transcript in each tumour we then used both the transcriptional orientation of the gene and the mutational asymmetry of the segment containing it to resolve the segregated lesions to either the template (anti-sense) or non-template (sense) strand of the gene. Transcripts contained in mutationally symmetric regions or not meeting the strict filtering criteria were excluded from analysis.

We then analysed mutation rates stratifying by gene expression level and the template/non-template strand of the lesions but aggregating between tumours within the same strain. The TSS coordinates used correspond to the annotated 5′ end of the primary transcripts.

Multiallelic variation

Aligned reads spanning genomic positions of somatic mutations were re-genotyped using Samtools mpileup (v.1.9)[55]. Genotypes supported by ≥ 2 reads with a nucleotide quality score of ≥ 20 were reported, considering sites with two alleles as biallelic, those with three or four alleles as multiallelic. The fraction of called mutations exhibiting multiallelic variation was calculated for the analysable fraction of the genome, across 10Mb consecutive windows and also for each of the mutational asymmetry segments calculated for each tumour.

A null expectation for the multiallelic rate estimate was generated per C3H tumour; genomic positions identified as mutated across the other 370 tumours were down-sampled to match the mutation count in the focal tumour. Any of these proxy mutation sites with a non-reference genotype supported by ≥ 2 reads and nucleotide quality score ≥ 20 at the focal site were referred to as 'multiallelic' for the purposes of defining a background expectation for the calling of multiallelic variation. For each tumour, this was repeated 100 times and the mean reported.

We used WES of 15 C3H tumours from prior work[6] that have subsequently been used to generate WGS data in this study as a basis for validating multiallelic calls. Multiallelic variant positions derived from WGS were genotyped in WES using Samtools mpileup, as described above. Only sites with ≥ 30× WES coverage were considered and alleles were found to be concordant if a WGS genotype was supported by ≥ 1 read in the WES data. To provide a null expectation, the analysis was repeated using WES data from a different tumour and validation rates reported for all versus all combinations of mismatched WGS–WES pairs ($n = 15^2 - 15 = 210$).

To quantify combinatorial genetic diversity for each tumour, pairs of mutations located between 3 and 150 nt apart were phased using sequencing reads that traversed both mutation sites. Distinct allelic combinations were counted after extraction with Samtools mpileup using only reads with nucleotide quality score ≥ 20 over both mutation sites.

Estimating the cell generation of transformation

Knowing the fraction of lesion segregation segments that generated multiallelic variation across a tumour genome allows the inference of the generation time post-mutagenesis of the cell from which the tumour developed, because each successive cell generation is expected to retain only 50% of the lesion containing segments. We estimate this fraction as follows. Let p denote the fraction of multiallelic segments and let q be its complement, that is, the fraction of non-multiallelic segments, for each tumour genome. Segment boundaries being SCE sites or chromosome boundaries. In order to determine p, we re-purpose the quadratic Hardy–Weinberg equation: $p + q = p^2 + 2pq + q^2 = 1$, which holds since the two possible fractions need to sum to unity. Given an asymmetric segment of interest in the diploid genome, there are three distinct scenarios: (i) both chromosomes are multiallelic (p^2), (ii) One of the chromosomes is multiallelic and the other is not ($pq + qp$) and (iii) both chromosomes are non-multiallelic (q^2). The first two scenarios are not distinguishable from the data as both appear multiallelic (m). However, in the third scenario, for a segment to be non-multiallelic (biallelic, b), both chromosomal copies have to be non-multiallelic. As described below, q^2 can be estimated directly from the data and is subsequently used to estimate $P = 1 - \sqrt{(q^2)}$ and hence the cell generation number of transformation post-mutagenesis.

The estimation of q^2 requires computing the ratio $q^2 = b/(b + m)$. We can directly observe the counts of b as non-multiallelic segments. The number of autosomal chromosome pairs ($n = 19$) and count of SCE events (x) give the total number of segments in the genome $b + m = n + x$. Exchange events are not expected to align between allelic chromosomes which will result in the partial overlap of segments between allelic copies. Although this increases the number of observed segments (b and m) relative to actual segments, assuming the independent behaviour of allelic chromosomes and that segment length is independent of multiallelic state, this partial overlap does not systematically distort the quantification of b or the estimation of q^2.

To call a non-multiallelic segment (b) we require less than 4% multiallelic sites. The threshold is based on the tri-modal frequency distribution of multiallelic rates per segment, aggregated over all 371 C3H tumours. The 4% threshold separates the lower distribution of multiallelic rates from the mid and higher distributions.

To test for the enrichment of specific driver gene mutations in early generation versus late generation transformation post-DEN treatment, we applied Fisher's exact test (fisher.test function in R) to compare the generation 1 ratio of tumours with, versus those without a focal mutation, to the same ratio for tumours inferred to have transformed in a later generation. We additionally report the same odds ratios, but requiring that the "with focal mutation" tumours had a driver mutation in only one of the driver genes: *Hras*, *Braf*, or *Egfr*.

Cell-line and human cancer mutation analysis

Somatic mutation calls were obtained from DNA maintenance and repair pathway perturbed human cells[26]. Of the 128,054 reported single nucleotide variants, 6,587 unique mutations (genomic site and specific change) were shared between two or more sister clones, so probably represent mutations present but not detected in the parental clone. All occurrences of the shared mutations were filtered out leaving 106,688 mutations for analysis, although the inclusion of these filtered mutations does not alter any conclusions drawn. Somatic mutation calls from mutagen exposed cells[5] were obtained, no additional filtering was applied to these sub-clone mutations.

Somatic mutation calls from the International Cancer Genome Consortium (ICGC)[56] were obtained as simple_somatic_mutation.open.* files from release 28 of the consortium, one file for each project. These somatic mutations have been called from a mixture of WGS and WES. Of the 18,965 patients represented (and not embargoed in the release 28 data set), 116 were excluded from analysis; these represent a distinct WES subset of the LICA-CN project that appear to show a processing artefact in the distribution of specific mutation subsets. ICGC mutations were filtered to remove insertion and deletion mutations and also filtered for redundancy so that each mutation was only reported once for each patient. Mutation signatures deconvolution was performed using the R MutationPatterns (v.1.4.2)[57] package and COSMIC signature 22 was interpreted as aristolochic acid[3].

The rl_{20} metric and runs tests

Amongst only the informative mutations (for example, T→N/A→N in DEN) three consecutive T→N without an intervening A→N is a run of three. The R function rle was used to encode the run-lengths for binary vectors of informative mutations along the genome of a focal tumour. Ranking them from the longest to the shortest run, we find the set of longest runs that encompass 20% of all informative mutations in the tumour. The run-length of the shortest of those is reported as the rl_{20} metric. The threshold percent of mutations was defined as having to be less than 50%, as on average only 50% of the autosomal genomes are expected to show mutational asymmetry patterns. On testing with randomized data, the value of 20% gave a stable null expectation (maximum observed value of a run of five when simulating 10,000 informative mutations) and still encompassed a large fraction of the informative mutations. All rl_{20} results reported were implemented so that runs were broken when crossing chromosome boundaries. To define an empirical significance threshold for genomes with fewer mutations, we simulated 1,000 random informative mutations 100,000 times, >99.995% simulations had $rl_{20} \leq 5$ and 100% $rl_{20} \leq 6$.

The Wald–Wolfowitz runs test was performed using the runs.test function of the R randtests (v.1.0)[58] library. It was applied to binary vectors of informative changes as described above, with threshold = 0.5.

The Wald–Wolfowitz runs test significance is inflated by coordinated dinucleotide changes, such as those produced by UV light exposure and also other local mutational asymmetries such as replication asymmetry[13] and kataegis events[14,59]. The rl_{20} metric appears robust to most such distortions but we find it efficiently detects kataegis events that are in an otherwise mutationally quiet background, as is often the case for breast cancer. For this reason we also indicate the total genomic span of mutations in the rl_{20} subset of mutation runs: kataegis events typically span a tiny (<5%) fraction of the whole genome.

Key resources

The key reagents and resources required to replicate our study are listed in Supplementary Table 3. For externally sourced data, where applicable, URLs that we used can be found in the Git repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

Primary data processing was performed in shell-scripted environments calling the software indicated. Except where otherwise noted, analysis processing post-variant calling was performed in a Conda environment and choreographed with Snakemake running in an LSF batch control system (Supplementary Table 3).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The WGS FASTQ files are available from the European Nucleotide Archive (ENA) under accession number PRJEB37808. RNA-seq files are available from Array Express under experiment number E-MTAB-8518. Digitised histology images are available from Biostudies under accession S-BSST383.

Code availability

The analysis pipeline including Conda and Snakemake configuration files can be obtained without restriction from the repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

35. Thoolen, B. et al. Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol. Pathol.* **38** (Suppl), 5S–81S (2010).
36. Lilue, J. et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Picard Tools (Broad Institute, 2019); http://broadinstitute.github.io/picard
39. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
40. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
41. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2013).
42. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLOS Comput. Biol.* **12**, e1004873 (2016).
43. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-seq quantification with kallisto. *Nat. Biotechnol.* **34**, 525–527 (2016).
44. Cunningham, F. et al. Ensembl 2019. *Nucleic Acids Res.* **47** (D1), D745–D751 (2019).
45. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
46. Church, D. M. et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
47. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
48. Armstrong, J. et al. Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. Preprint at *bioRxiv* https://doi.org/10.1101/730531 (2019).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at *bioRxiv* https://doi.org/10.1101/372896 (2018).
51. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
52. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 5396 (2019).
53. Killick, R. & Eckley, I. A. changepoint: an R package for changepoint analysis. *J. Stat. Softw.* **58**, 1–19 (2014).
54. Akeson, E. C. et al. Chromosomal inversion discovered in C3H/HeJ mice. *Genomics* **87**, 311–313 (2006).
55. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
57. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
58. Caeiro, F. & Mateus, A. randtests: testing randomness in R. (2014).
59. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
60. Singer, B. In vivo formation and persistence of modified nucleosides resulting from alkylating agents. *Environ. Health Perspect.* **62**, 41–48 (1985).

Acknowledgements We thank M. Roller and F. Markowetz for supervision; L. Mularoni and G. Ritchie for software support; the CRUK Cambridge Institute Core facilities for their valuable contribution: CRUK Biological Resources (A. Mowbray), Preclinical Genome Editing (L. Young, S. Kupczak, M. Cronshaw, P. Mackin, Y. Cheng and L. Hughes-Hallett), Genomics (J. Hadfield and F. Bowater), Bioinformatics (G. Brown, M. Eldridge and R. Bowers), Histopathology and ISH (L.-A. McDuffus, C. Brodie and J. Arnold), and Research Instrumentation (J. Gray); Edinburgh Genomics (Clinical) Facility; the EMBL-EBI technical services cluster (Z. Mears, A. Cristofori, T. Nowak, S. Nanuwa, V. Tabak and A. Checcucci); and W. Bickmore and C. Ponting for comments on the manuscript. This work was supported by: Cancer Research UK (20412 and 22398), the European Research Council (615584, 682398), the Wellcome Trust (WT108749/Z/15/Z, WT106563/Z/14/A and WT202878/B/16/Z), the European Molecular Biology Laboratory, the MRC Human Genetics Unit core funding programme grants (MC_UU_00007/11 and MC_UU_00007/16) and the ERDF/Spanish Ministry of Science, Innovation and Universities-Spanish State Research Agency/DamReMap Project (RTI2018-094095-B-I00). S.J.A. received a Wellcome Trust PhD Training Fellowship for Clinicians (WT106563/Z/14/Z) and is now funded by a National Institute for Health Research (NIHR) Clinical Lectureship. O.P. is funded by a BIST PhD fellowship supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the Barcelona Institute of Science and Technology. V.S. is supported by an EMBL Interdisciplinary Postdoc (EIPOD) fellowship under Marie Skłodowska Curie actions COFUND (664726). E.K. is supported by the EMBL International PhD Programme. C.A-P. is supported by La Caixa Foundation fellowship (ID 100010434; LCF/BQ/ES18/11670011). S.V.B. is supported by ERC Starter Grant 759967. A.E. is supported by a UKRI Innovation Fellowship (MR/RO26017/1). A.K. is a cross-disciplinary postdoctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (core grant to the MRC Institute of Genetics and Molecular Medicine). I.S. is supported by an FPI fellowship from Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from Spanish Ministry of Science, Innovation and Universities (MICINN, Government of Spain) and is supported by CERCA (Generalitat de Catalunya).

Author contributions S.J.A., F.C., C.F. and D.T.O. conceived the project and designed the experiments. S.J.A., F.C. and C.F., performed the mutagenesis experiments and sequencing experiments. E.L.-A and A.M.R. performed supporting experiments. J.S.-L. provided contract sequencing. S.J.A. performed the histopathological analyses with S.E.D. providing advice. C.J.A. and M.S.T. designed and implemented computational analysis. M.S.T. discovered lesion segregation. O.P., V.S., T.F.R., M.L., S.A., E.K. and J.L. performed supporting computational analysis. C.A-P., S.V.B., R.M.D., A.E., V.B.K., A.K., I.S. and L.T. contributed to the computational analyses. T.F.R., M.L., S.A. and A.D.Y. curated data. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. supervised the work. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. lead the Liver Cancer Evolution Consortium. S.J.A. and P.F. provided scientific and administrative organisation. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. funded the work. S.J.A., D.T.O. and M.S.T. wrote the manuscript. All authors had the opportunity to edit the manuscript. All authors approved the final manuscript.

Competing interests P.F. is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd. The other authors declare no competing interests.

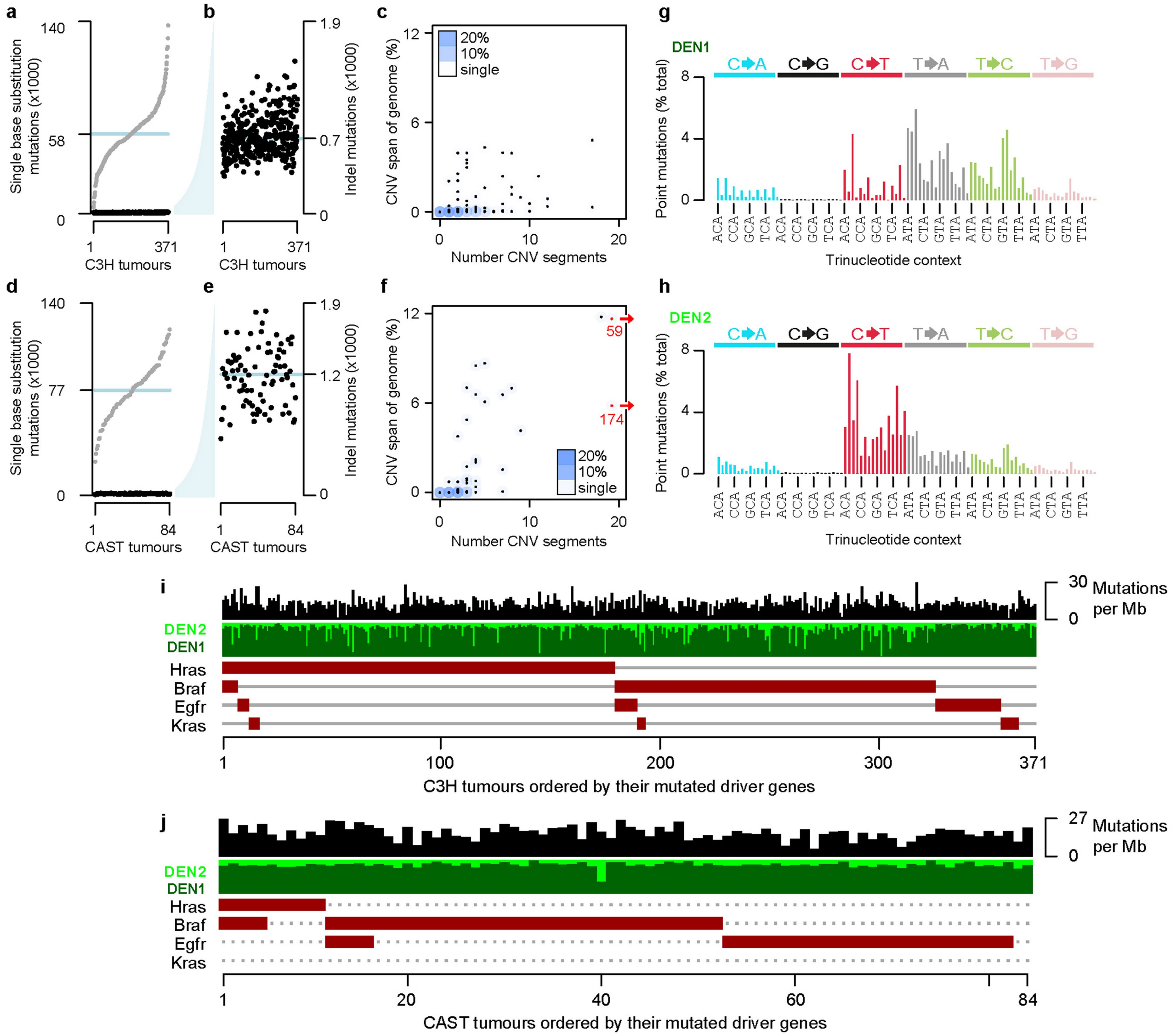
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2435-1.

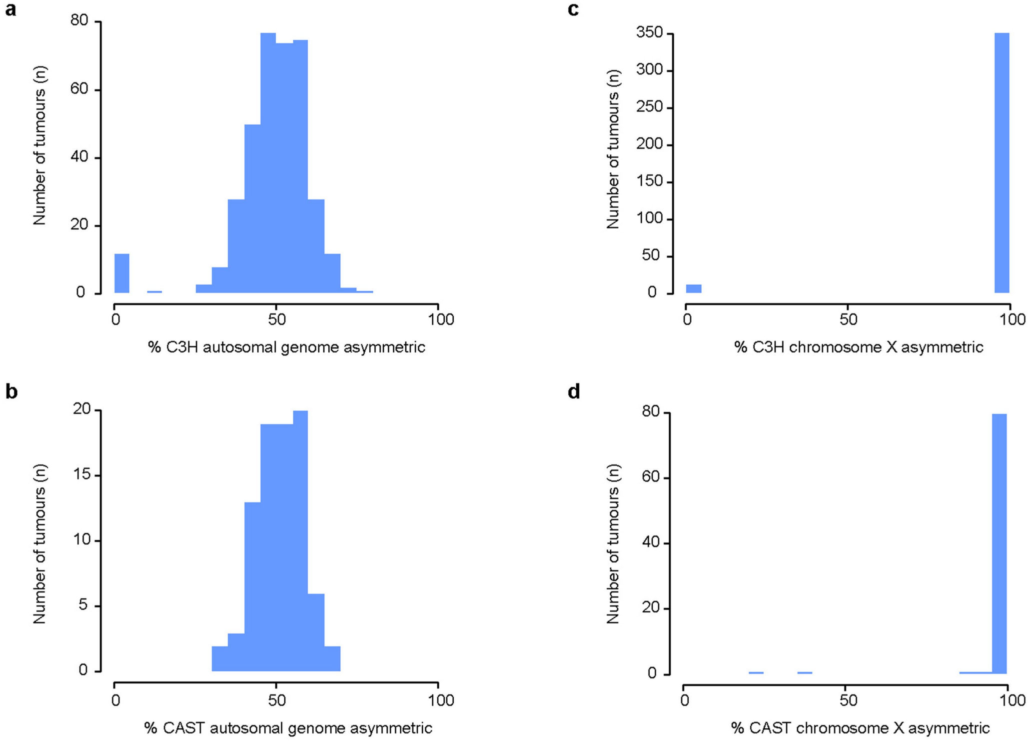
Correspondence and requests for materials should be addressed to D.T.O. or M.S.T.

Peer review information *Nature* thanks Trevor Graham and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

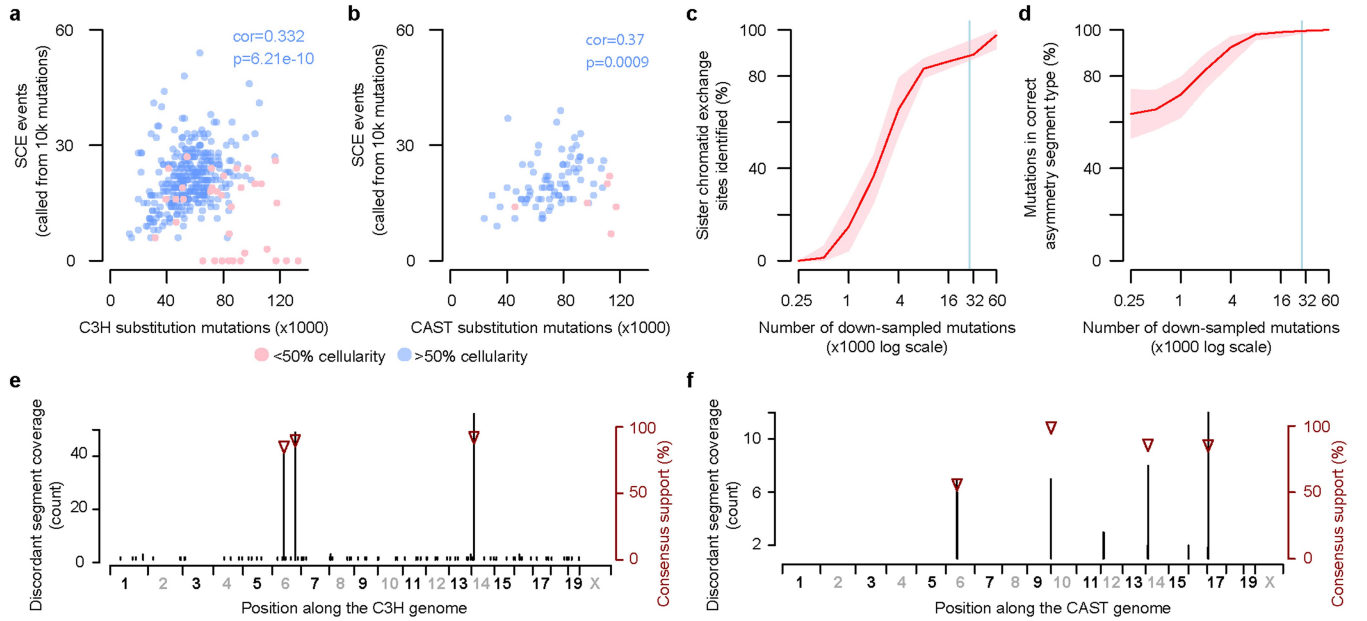
Reprints and permissions information is available at http://www.nature.com/reprints.



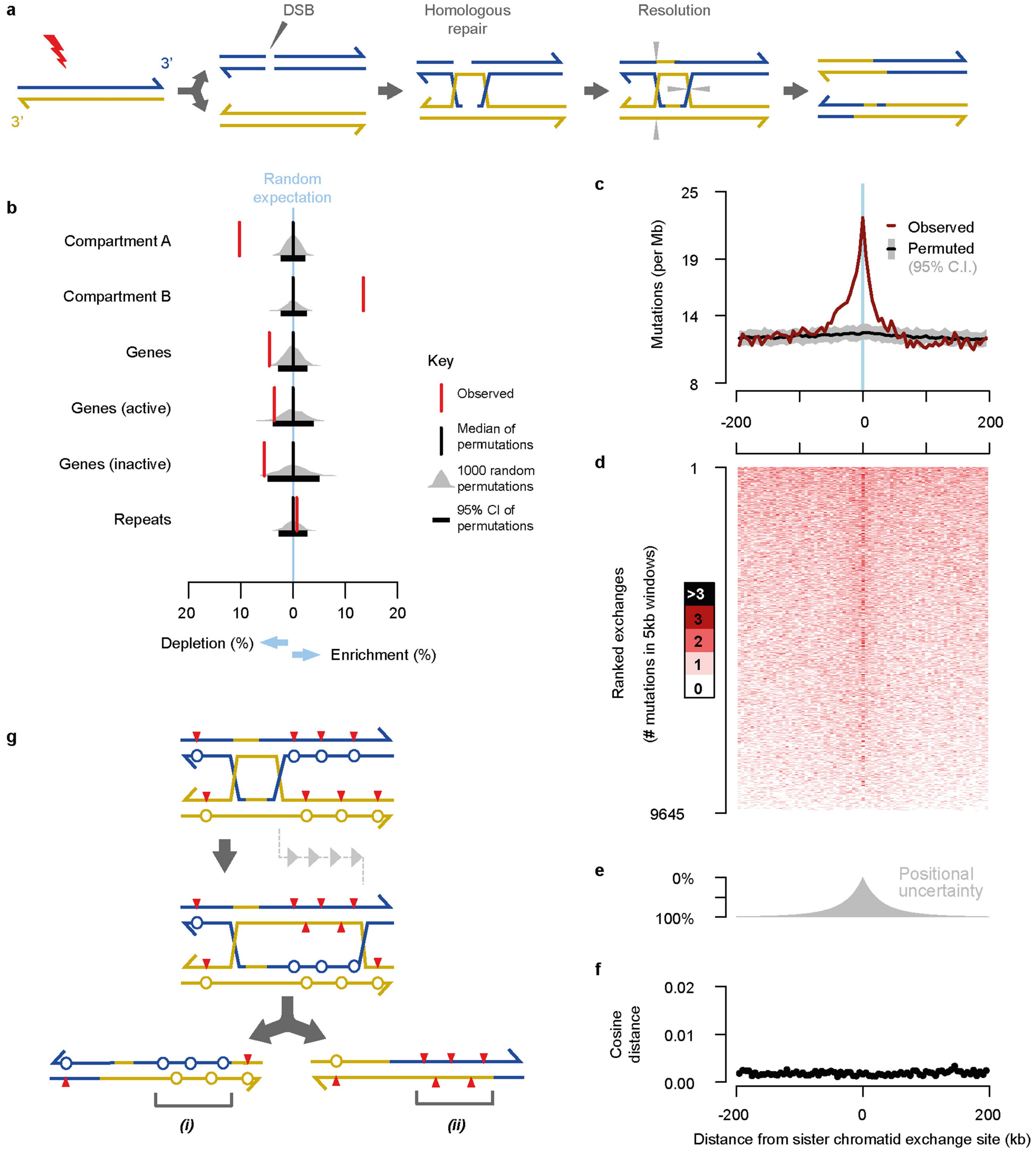
Extended Data Fig. 1 | Summary mutation metrics for C3H and CAST tumours. a, Single nucleotide substitution rates per C3H tumour, rank ordered over *x*-axis (grey points, median blue line). Insertion/deletion (indel, <11 nt) rates show as black. **b**, Y-axis from **a**, expanded to show distribution of indel rates with preserved tumour order. **c**, Number of C3H copy number variant (CNV) segments and their total span as a percent of the haploid genome. Blue shading shows intensity of overlapping points as a percent of all tumours in the plot. **d**–**f**, Corresponding plots for CAST derived tumours; **f**, two extreme *x*-axis outliers relocated (red) and *x*-axis value shown. **g**, **h**, Mutation spectra deconvolved from the aggregate spectra of 371 C3H tumours, subsequently referred to as the DEN1 and DEN2 signatures. DEN1 is dominated by T→N or their complement A→N changes thought to arise from the O^4-ethyl-deoxythymidine adduct of T[10]. DEN2 substitutions are primarily C→T or their complement G→A changes likely from O^6-ethyl-2-deoxyguanosine lesions of G[10]. **i**, Oncoplot summarizing mutation load, mutation signature composition, and driver gene mutation complement of C3H tumours. **j**, Oncoplot of CAST derived tumours as in **i**. The DEN2 signature is a minor component of most tumours but prominent in a minority (**i**, **j**).



Extended Data Fig. 2 | Mutational asymmetry across 50% of the autosomal genome and 100% of the haploid X chromosomes. **a**, **b**, Typically 50% of the autosomal genomic span (percent of nucleotides) in tumours is contained in segments with either Watson or Crick strand mutational asymmetry. **a**, C3H tumours, *n* = 371. **b**, CAST tumours, *n* = 84. **c**, **d**, Typically 100% of the haploid X chromosome shows Watson or Crick strand mutational asymmetry. **c**, C3H tumours (*n* = 371). **d**, CAST tumours (*n* = 84).

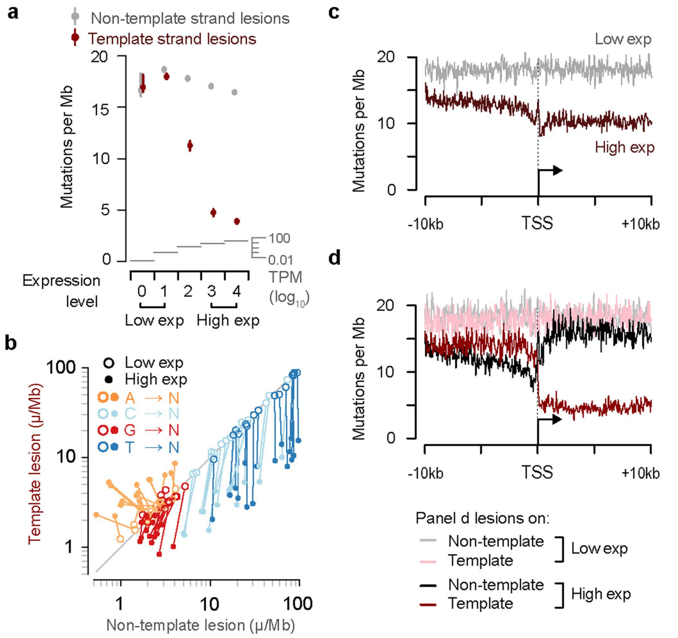


Extended Data Fig. 3 | The frequency of SCEs correlates with mutation rate, and localizing reference genome assembly errors. a, The relationship between single nucleotide substitution mutation load and detected SCE events in C3H tumours. DEN is known to produce ethyl adducts on the sugar-phosphate backbone of DNA as well as mutation-inducing modifications to the bases[10] which could lead to strand breaks[60] triggering SCE. The frequent observation and correlation between rates of SCE and point mutation supports this view. Counts of SCE (*y*-axis) are based on down-sampling to 10,000 informative mutations per tumour to ensure equal power to detect SCE in each tumour. Tumours with <50% cellularity (pink) have high mutation load and form a sub-group with few detected SCE events; these are suspected to be polyclonal tumours and were excluded from the Pearson's correlation reported ($n = 335$ independent tumour samples, implemented in a two-sided test, significance from Fisher's transform). **b**, As for **a**, but showing CAST derived tumours ($n = 84$, after cellularity exclusions $n = 77$). **c**, Evaluation of the relationship between mutation load and ability to detect SCE events. Mutations from C3H tumour 94315_N8 (shown in Fig. 2) randomly down-sampled and segmentation analysis applied. The *y*-axis shows the percentage of SCE events detected (100 replicates, mean red, 95% C.I. pink). The *x*-axis is on a log-scale: 95% of C3H and >95% of CAST tumours have mutation counts to the right of the blue vertical line. Down-sampling other tumours gave comparable results. **d**, The same down-sampling data as shown in **c** but the *y*-axis shows the percent of mutations with the correct (same as full data) mutational asymmetry assignment (mean red, 95% C.I. pink). **e**, Candidate C3H reference genome assembly errors. Genome coordinates shown on the *x*-axis. Immediate switches between Watson and Crick asymmetry are not expected on autosomes unless both copies of the chromosome have a SCE event at equivalent sites. However, inversions and translocations between the sequenced genomes and the reference assembly are expected to produce immediate asymmetry switches. The discordant segment coverage count (black *y*-axis) shows the number of informative tumours (those with either Watson or Crick strand asymmetry at the corresponding genome position) that suggest a tumour genome to reference genome discrepancy. Consensus support (brown *y*-axis) plotted as triangles shows the percentage of informative tumours that support a genomic discrepancy at the indicated position (only shown for values >50% support). The two sites on chromosome 6 in C3H correspond to a previously identified C3H strain specific inversion that is known to be incorrectly oriented in the C3H reference assembly[54]. **f**, Candidate CAST reference genome assembly errors, plotted as per **e**. The candidate misassembly on chromosome 14 in both strains occurs at an approximately orthologous position, suggesting a rearrangement shared between strains or a misassembly in the BL6 GRCm38 reference assembly against which other mouse reference genome assemblies have been scaffolded.

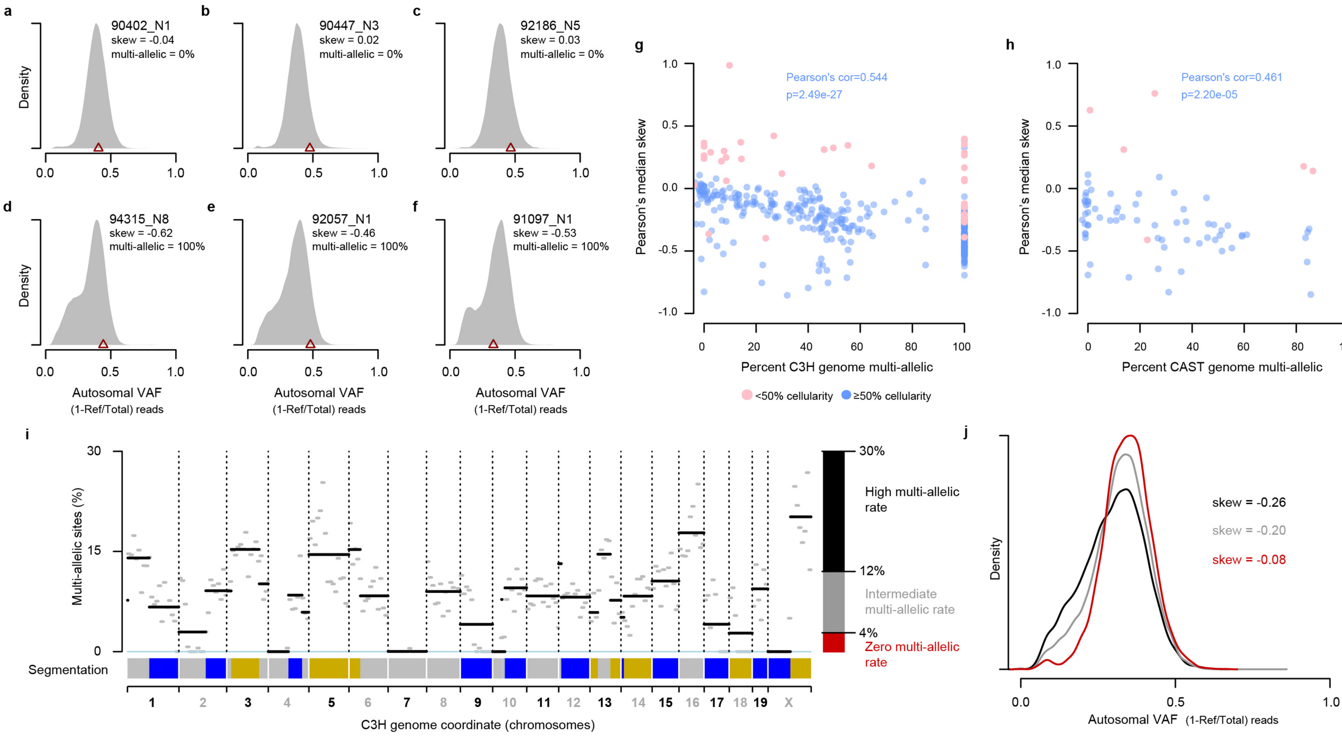


Extended Data Fig. 4 | See next page for caption.

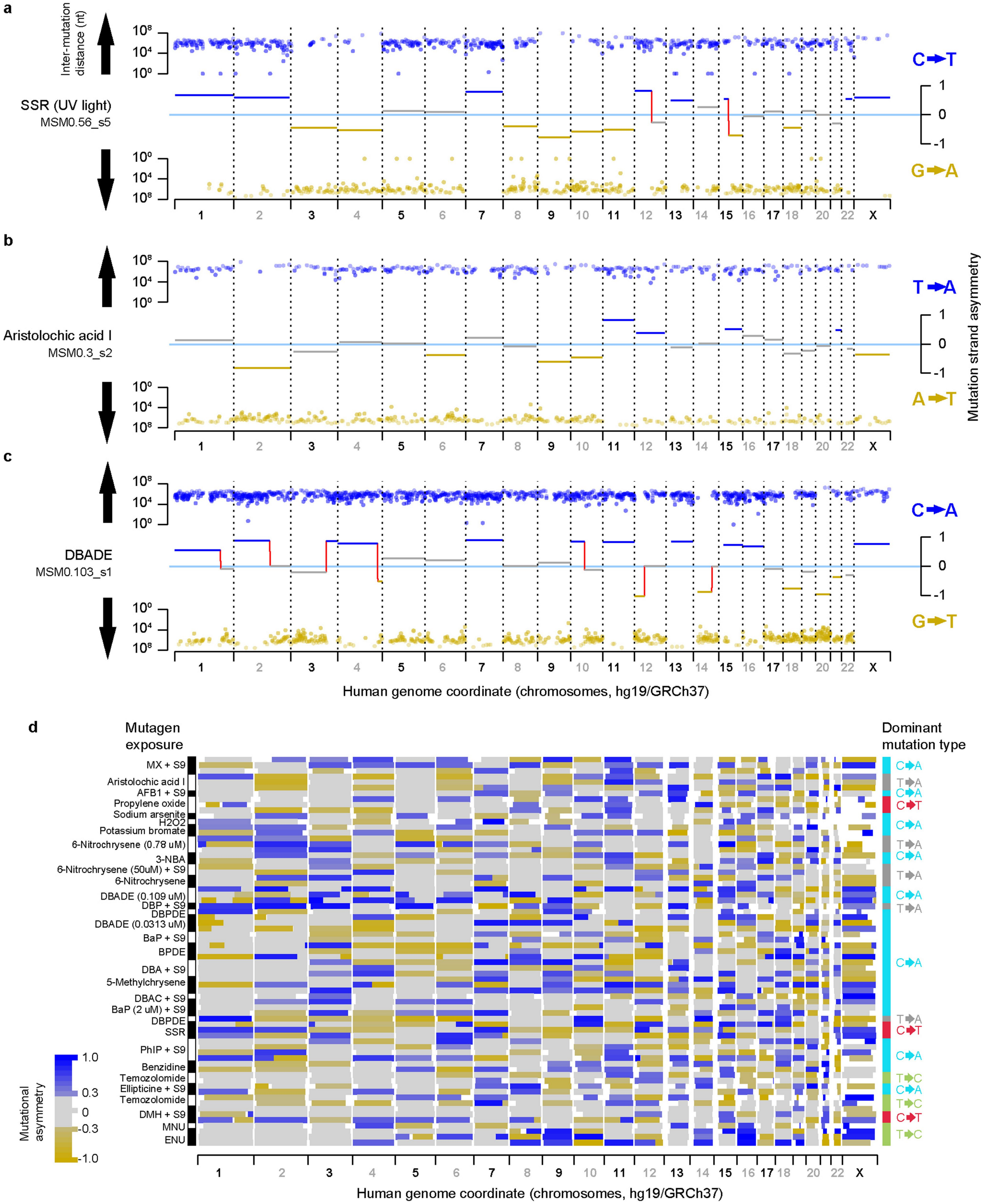
Extended Data Fig. 4 | Locally elevated mutation load is driven by SCE. **a**, Double strand breaks (DSBs) and other DNA damage can trigger homologous-recombination-mediated DNA repair between sister chromatids. The repair intermediate resolves into separate chromatids through cleavage and ligation; grey triangles denote cleavage sites for one of the possible resolutions that would result in a large-scale SCE event. Although illustrated for double-ended DNA breaks, single ended breaks from collapsed replication forks can be repaired through homologous recombination and could similarly lead to the formation of repair intermediate structures that can be resolved as SCEs. **b**, Enrichment analysis of SCE sites (red) compared with null expectations from randomly permuting locations into the analysable fraction of the genome (grey distributions), the black boxes denote 95% of 1,000 permutations. SCE events are enriched in later replicating and transcriptionally less active genomic regions (Hi-C defined compartment B), and correspondingly depleted from early replicating active regions. **c**, Aggregating across $n = 9{,}645$ SCE sites, the observed mutation rate approximately doubles at the inferred site of exchange ($x = 0$). Aggregate mutation rates (brown) were calculated in consecutive 5-kb windows. Compositionally matched null expectation was generated by permuting each exchange site into 100 proxy tumours and calculating median (black) and 95% confidence intervals (grey) while preserving the total number of projected sites per proxy tumour. **d**, The elevated mutation count is not the result of a high mutation density in a subset of exchange sites, rather it is a subtle increase in mutations across most exchange sites. Heatmap showing mutation counts calculated in consecutive 5-kb windows across each exchange site. Rows represent each exchange site, rank-ordered by total mutation count across each 400-kb interval. **e**, The distribution of positional uncertainty in exchange site location approximately mirrors the decay profile of elevated mutation frequency. **f**, Divergence of mutation rate spectra is shown as cosine distance between the analysed window and the genome wide mutation rate spectrum aggregated over all C3H tumours. Despite the elevated mutation frequency, there is no detected distortion of the mutation spectrum. **g**, A model based on homologous recombination repair intermediate, branch migration that produces heteroduplex segments of (i) mismatch:mismatch (circles) and (ii) lesion:lesion (red triangles) strands. Subsequent strand segregation would increase the mutational diversity of a descendant cell population but not the mutation count per cell (key as per Fig. 2).



Extended Data Fig. 5 | Replication of TCR with lesion strand resolution in *Mus musculus castaneus*. **a**, TCR of template strand lesions is dependent on transcription level (P15 liver, median TPM). Mutation rate estimates (circles) are the aggregate rates for expression level binned genes across CAST tumours (*n* = 84). Expression level bin 0 contains *n* = 2,645 genes, all subsequent bins contain *n* = 4,323 genes. See Methods for per-gene, per-tumour inclusion criteria. Empiric confidence intervals (99%) were calculated through bootstrap sampling (*n* = 100 replicates) of genes within the expression level bin. **b**, Comparison of mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **c**, Sequence composition normalized profiles of mutation rate around TSS loci. **d**, Stratifying the data plotted in **c** by lesion strand reveals greater detail on the observed mutation patterns, including the pronounced influence of bidirectional transcription initiation.

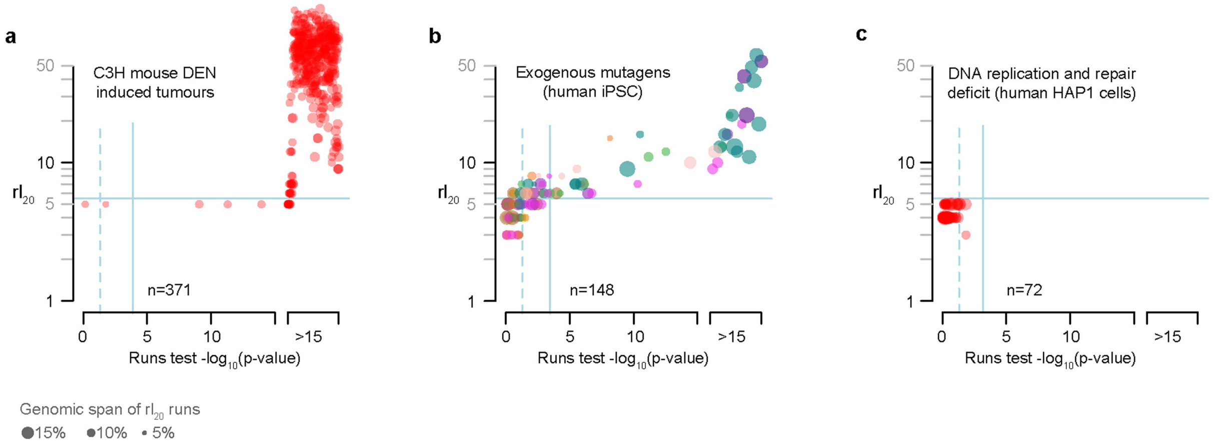


Extended Data Fig. 6 | Variant allele frequency distributions demonstrate high rates of non-mutagenic replication over segregating lesions. a–**f**, VAF distributions shown as probability density functions (total area under curve = 1) for six example tumours, calculated taking into account observed multiallelic variation. The VAF for identified driver mutations is indicated (brown triangle). Tumour identifiers are shown top right along with the percent of genomic segments (based on mutation asymmetry segmentation) that are multiallelic. Skew shows Pearson's median skewness coefficient for the VAF distributions. **a**–**c**, Tumours with no multiallelic segments and exhibit a symmetric VAF distribution showing minimal sub-clonal structure. **d**–**f**, Tumours with all segments multiallelic, illustrating the sub-clonal structure generated by segregating lesions. **g**, Tumours with a high proportion of multiallelic segments have a left-skewed VAF distribution indicating frequent non-mutagenic replication over segregating lesions. Percent of genome segments that are multiallelic (*x*-axis) plotted against VAF distribution skew for 371 C3H tumours. Tumours with low estimated cellularity indicated in pink and excluded from correlation analysis ($n = 335$ independent tumour samples in Pearson's correlation, two-sided significance from Fisher's transform). **h**, As for **g**, but showing 84 CAST tumours ($n = 77$ independent tumours included in Pearson's correlation). **i**, Mutation asymmetry summary ribbon for example C3H tumour 90797_N2; C3H genome on the *x*-axis. The percent of mutation sites with robust support for multiallelic variation (*y*-axis) calculated in 10 Mb windows (grey) and for each asymmetric segment (black). Thresholds for high (black), intermediate (grey) and zero (red) rates of multiallelic sites shown on the right axis. **j**, VAF density plots for the example tumour 90797_N2 (shown in **i**) mutations in asymmetry segments stratified by the multiallelic rate thresholds defined in **i**. As with individual tumour based analysis (**a**–**h**), high multiallelic rates correspond to a leftward skew of the VAF (black, grey) whereas segments without multiallelic variation (red) show a minimally skewed distribution.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Examples of mutation patterns generated by lesion segregation from a diverse range of clinically relevant mutagens. **a**–**c**, Genome-wide mutation asymmetry plots (shown as per Fig. 2a–c) for mutagen exposed human iPSCs[5]. Cells exposed to simulated solar radiation illustrate lesion segregation for ultraviolet damage (**a**). Immediately adjacent mutations (intermutation distance 10^0) indicate CC→TT dinucleotide changes. Despite a low total mutation load (1,308 nucleotide substitutions, 842 informative T→A changes), the mutational asymmetry of lesion segregation is evident for the aristolochic acid exposed clone[5] (**b**) and the polycyclic aromatic hydrocarbon DBADE (**c**) that is found in tobacco smoke. **d**, Summary mutation asymmetry ribbons (as per Fig. 2d) for all mutagen exposed clones with $rl_{20} > 5$, which illustrates the independence of asymmetry pattern between replicate clones, almost universal asymmetry on chromosome X, and approximately 50% of the autosomal genome with asymmetry over autosomal chromosomes. The dominant mutation type is indicated for each mutagen. In those clones with low mutation rates, some sister exchange sites are likely to have been missed leading to reduced asymmetry signal (for example, on the X chromosome). Segments with <20 informative mutations are shown in white.



Extended Data Fig. 8 | Lesion segregation is evident for multiple DNA damaging agents but not for damage independent mutational processes. **a**, DEN induced C3H tumour genomes ($n = 371$) typically show significant mutational asymmetry across their genome. Wald–Wolfowitz runs test (x-axis) P-values calculated using a normal approximation (two-sided). Nominal $P = 0.05$ significance threshold indicated by dashed blue line, Bonferroni-corrected threshold shown as solid vertical blue line. P-values $<1 \times 10^{-15}$ are rank-ordered. The rl$_{20}$ metric (Fig. 5a; Methods) is shown on the y-axis, horizontal blue line gives emprical significance threshold of rl$_{20} > 5$. **b**, Many human iPSCs grown from single cells after exogenous mutagen exposure[5] show significant mutation asymmetry ($n = 148$ WGS, mutagen-exposed cell lines). Statistical calculations and plotting as in **a**, with adjustment of Bonferroni correction. Diverse categories of mutagen, denoted by point colour (see Fig. 5b), show asymmetry indicative of lesion segregation. **c**, Cell lines with genetically perturbed genome replication and maintenance machinery[26] and similar mutation load to those in **b** do not show significant mutation asymmetry ($n = 72$ WGS, genetically perturbed cell-lines). Statistical calculations and plotting as in **a** with adjustment of Bonferroni correction.

Extended Data Table 1 | A lesion segregation-based test for oncogenic selection

Strain	Gene	Mutation	Mutation count	Odds ratio	P-value	Known driver
C3H	Braf	6:37548568_A/T	151	2.13	5.77×10^{-6}	Yes
C3H	Hras	7:145859242_T/C	81	2.67	6.88×10^{-6}	Yes
C3H	Hras	7:145859242_T/A	65	1.02	1	Yes
C3H	Intronic Fmnl1	11:105081902_A/C	44	1.03	1	No
C3H	Intergenic	9:73125689_G/C	42	1.13	1	No
C3H	Egfr	11:14185624_T/A	34	3.87	1.23×10^{-4}	Yes
CAST	Braf	6:37451282_A/T	42	1.41	0.338	Yes

Recurrently mutated sites in C3H and CAST tumours with sufficient estimated power to detect oncogenic selection through biased strand retention analysis (required >33 C3H recurrences or >41 CAST recurrences). Odds ratio values >1 indicate the predicted correlation of driver mutation and Watson/Crick strand retention in tumours with the candidate driver mutation, but not for those without the mutation. The Fisher's exact test was performed on counts of chromosomes with Watson and Crick strand asymmetries (Methods). Each tested site was autosomal, thus total sample sizes were: $n = 2 \times 371 = 742$ for C3H, and $n = 2 \times 84 = 168$ for CAST. *P*-values (two-sided) are shown after Bonferroni correction (7 tests performed). Known driver indicates the mutation or its orthologous change has previously been implicated as a driver of hepatocellular carcinoma[6]. The CAST 6:37451282_A/T mutation is orthologous to the C3H 6:37548568_A/T mutation.

nature research

Corresponding author(s): Martin Taylor
Duncan Odom

Last updated by author(s): Apr 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina Software Control (ICS) 3.3.76
Data analysis	Open source data analysis tools were used, all documented with version numbers in the manuscript and Supplemental Table 3. The data analysis code specific to this project is available at https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls Software and versions used in the analysis: bwa mem 0.7.12 CNVkit 0.9.6 kalisto 0.43.1 strelka2 2.8.4 Bcftools 1.9 Cactus toolkit 2.1 Ensembl VEP 91.3 Juicebox 7.5 oncodriveCLUSTL 1.1.1 onocdriveFML 2.2.0 picard 1.124 Python 3.6 R 3.5.1 r-bit64 0.9_7 r-cairo 1.5_9 r-changepoint 2.2.2 r-colorbrewer 1.1_2

r-data.table 1.11.4
r-hashmap 0.2.2
r-randtests 1.0.0
r-sigFit 2.0.0
r-zoo 1.8_3
Samtools 1.9
snakemake 5.2.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The WGS FASTQ files are available from the European Nucleotide Archive (ENA) under accession: PRJEB37808. RNA-seq files are available from Array Express E-MTAB-8518. Digitised histology images are available from Biostudies under accession S-BSST383.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for the main DEN-induced oncogenesis study were chosen to be the maximum possible with available resource, and included whole genome sequencing of more tumours that the pilot exome-sequencing project (Connor et al, J Hepatol. 2018). This earlier exome-only study identified mutation spectra and driver mutations in known cancer genes; this ensured that the larger whole genome study, including replication in a second strain, was sufficiently powered to detect cancer genome properties of interest. For some statistical tests (e.g. oncogenic selection) a simulation study was performed to calculate thresholds for 80% power prior to performing analysis. As documented in the manuscript.
Data exclusions	116 exome sequenced human tumours from the ICGC LICA-CN project were excluded as documented in the manuscript methods section and Supplemental Table 2. These appear to suffer from a data processing artifact in the reported mutation calls (some types of mutation call are missing) and lower level data is not currently available. No exclusion criteria were pre-established.
Replication	Replication in a second mouse strain. Replication of mutation calls and multi-allelic variation in deeply sequenced exomes to confirm whole genome sequencing results (replication successful). Replication in human mutagen exposed cell lines (replication successful). Replication in publicly available human cancer data (replication successful).
Randomization	Multiple permutation based randomisations were performed, all documented in the manuscript and in the available code.
Blinding	Haematoxylin and eosin stained sections of liver tumours were blinded and assessed twice by a pathologist and discordant results reviewed by an independent hepatobiliary pathologist. Blinding was not relevant for genomic analyses as all processing was automated and all sequenced tumours were processed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	15-day-old (P15) male C3H and CAST mice were treated with a single intraperitoneal (IP) injection of N-Nitrosodiethylamine (DEN; Sigma-Aldrich N0258; 20 mg/kg body weight) diluted in 0.85% saline. Liver tumour samples were collected from DEN-treated mice 25 weeks (C3H) or 38 weeks (CAST) after treatment.
Wild animals	This study did not use wild animals.
Field-collected samples	This study did not use field-collected samples.
Ethics oversight	Animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical Review Body (AWERB).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Extensive signal integration by the phytohormone protein network

https://doi.org/10.1038/s41586-020-2460-0

Received: 6 June 2019

Accepted: 14 April 2020

Melina Altmann[1,12], Stefan Altmann[1,12], Patricia A. Rodriguez[1], Benjamin Weller[1], Lena Elorduy Vergara[1], Julius Palme[1,10], Nora Marín-de la Rosa[1], Mayra Sauer[1], Marion Wenig[2], José Antonio Villaécija-Aguilar[3], Jennifer Sales[2], Chung-Wen Lin[1], Ramakrishnan Pandiarajan[1], Veronika Young[1], Alexandra Strobel[1], Lisa Gross[4], Samy Carbonnel[5], Karl G. Kugler[6], Antoni Garcia-Molina[1,11], George W. Bassel[7], Claudia Falter[1], Klaus F. X. Mayer[6,8], Caroline Gutjahr[3,5], A. Corina Vlot[2], Erwin Grill[4] & Pascal Falter-Braun[1,9] ✉

Plant hormones coordinate responses to environmental cues with developmental programs[1], and are fundamental for stress resilience and agronomic yield[2]. The core signalling pathways underlying the effects of phytohormones have been elucidated by genetic screens and hypothesis-driven approaches, and extended by interactome studies of select pathways[3]. However, fundamental questions remain about how information from different pathways is integrated. Genetically, most phenotypes seem to be regulated by several hormones, but transcriptional profiling suggests that hormones trigger largely exclusive transcriptional programs[4]. We hypothesized that protein–protein interactions have an important role in phytohormone signal integration. Here, we experimentally generated a systems-level map of the *Arabidopsis* phytohormone signalling network, consisting of more than 2,000 binary protein–protein interactions. In the highly interconnected network, we identify pathway communities and hundreds of previously unknown pathway contacts that represent potential points of crosstalk. Functional validation of candidates in seven hormone pathways reveals new functions for 74% of tested proteins in 84% of candidate interactions, and indicates that a large majority of signalling proteins function pleiotropically in several pathways. Moreover, we identify several hundred largely small-molecule-dependent interactions of hormone receptors. Comparison with previous reports suggests that noncanonical and nontranscription-mediated receptor signalling is more common than hitherto appreciated.

To examine phytohormone signal integration by the plant protein network, we first identified 1,252 genes with probable or genetically demonstrated functions in phytohormone signalling (Fig. 1a and Supplementary Table 1). The corresponding network of literature curated binary interactions (LCIs) from the IntAct database[5] (LCI_{IntA}) shows extensive intrapathway but sparse interpathway connectivity (Extended Data Fig. 1), which could reflect an insulated organization of hormone signalling, or could be an artefact of inspection biases[6]. We therefore experimentally generated a systematic (unbiased design) map of the phytohormone signalling network. After cloning open reading frames (ORFs) for 1,226 (98%) of the selected genes (PhyHormORFeome), fivefold investigation of the pairwise matrix using a high-quality yeast two-hybrid (Y2H)-based mapping pipeline[7] yielded the phytohormone interactome main (PhI_{MAIN}) network. To find links into the broader *Arabidopsis* network, we screened PhyHormORFeome against roughly 13,000 *Arabidopsis* ORFs[8], resulting in an asymmetric PhI_{EXT} data set. We also conducted focused screens for interactions of pathway-specific repressors with transcription factors[9]($PhI_{Rep\text{-}TF}$), and for hormone-dependent interaction partners of phytohormone receptor proteins (PhI_{HORM}). In the stringent final step of the common Y2H pipeline, all candidate pairs were fourfold-verified (Fig. 1b).

The combined PhI network contains 2,072 interactions, of which 1,572 were previously unknown (Fig. 1c, Extended Data Fig. 1 and Supplementary Table 2). The interaction density in the symmetrically investigated PhI_{MAIN} (0.4‰) is higher than in the proteome-scale Arabidopsis Interactome-1 (AI-1, 0.1‰)[10], but lower than in the abscisic-acid (ABA)-focused interactome (7.5‰)[3]. It is likely that the increasing focus on functionally coherent proteins underlies this trend, but system

---

[1]Institute of Network Biology (INET), Helmholtz Center Munich, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [2]Inducible Resistance Signaling Group, Institute of Biochemical Plant Pathology (BIOP), Helmholtz Center Munich, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [3]Plant Genetics, TUM School of Life Sciences, Technical University of Munich (TUM), Freising, Germany. [4]Botany, TUM School of Life Sciences, Technical University of Munich (TUM), Freising, Germany. [5]Genetics, Faculty of Biology, Ludwig-Maximilians-Universität (LMU) München, Planegg-Martinsried, Germany. [6]Plant Genome and Systems Biology (PGSB), Helmholtz Center Munich, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [7]School of Biosciences, University of Birmingham, Birmingham, UK. [8]TUM School of Life Sciences, Technical University of Munich (TUM), Freising, Germany. [9]Microbe-Host Interactions, Faculty of Biology, Ludwig-Maximilians-Universität (LMU) München, Planegg-Martinsried, Germany. [10]Present address: Department of Genetics, Harvard Medical School, Boston, MA, USA. [11]Present address: Plant Molecular Biology, Faculty of Biology, Ludwig-Maximilians-Universität (LMU) München, Planegg-Martinsried, Germany. [12]These authors contributed equally: Melina Altmann, Stefan Altmann. ✉e-mail: pascal.falter-braun@helmholtz-muenchen.de

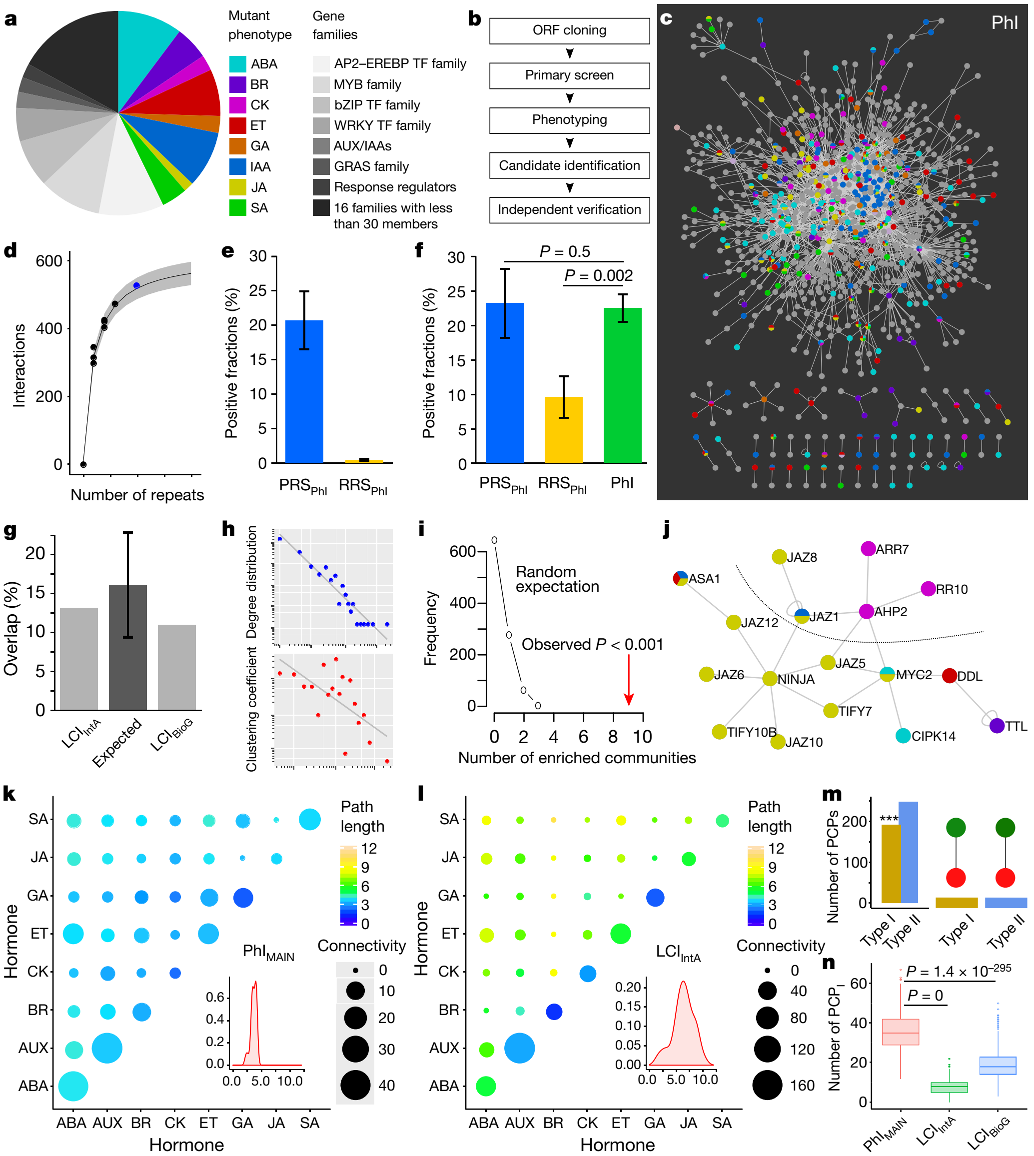


Fig. 1 | Phytohormone network mapping and analysis. a, Mutant phenotypes and overrepresented families in PhyHormORFeome candidates. ABA, abscisic acid; AUX, auxins; BR, brassinosteroids; CK, cytokinins; ET, ethylene; GA, gibberellic acids; IAA, indole-3-acetic acid; JA, jasmonic acid; KAR, karrikins; SA, salicylic acid; TF, transcription factor. **b**, Pipeline for mapping protein interactions. **c**, Phytohormone interactome (PhI) network. For node colours, see **a**. **d**, PhI$_{MAIN}$ sampling sensitivity: black dots show the verified interactions of the first three primary-screen repeats ($n = 3$); the black line shows the screen saturation model based on three repeats (with the grey corridor representing the standard error); blue dots show identified interactions after five repeats. **e**, Y2H assay sensitivity: positive fractions of PRS$_{PhI}$ ($n = 92$) and RRS$_{PhI}$ ($n = 95$). Error bars represent the standard error. **f**, Validation results: positive fractions of PRS$_{PhI}$ ($n = 69$), RRS$_{PhI}$ ($n = 83$) and PhI ($n = 285$). One-sided Fisher's exact test; error bars represent the standard error of proportion. **g**, Overlap of PhI$_{MAIN}$ with LCI$_{IntA}$ ($n = 109$) and LCI$_{BioG}$ ($n = 150$ interactions). Error bars show the propagated standard error. **h**, Distribution of the PhI$_{MAIN}$ degree and clustering coefficient. **i**, Number of hormone-signalling-function-enriched communities in PhI$_{MAIN}$ (red arrow) compared with $n = 1{,}000$ randomized control networks. **j**, JA- and CK-enriched community links (for node colours, see **a**). **k**, **l**, Distances between pathway combinations in PhI$_{MAIN}$ (**k**) and LCI$_{IntA}$ (**l**). Colours show average shortest distances; circle sizes show connection counts; insets show shortest distance distributions. **m**, Count of type I PCPs ($n = 192$) and type II PCPs ($n = 248$) in PhI$_{MAIN}$; *P* value is from the analysis in **n** (*** $P \leq 0.001$). **n**, Proportion of PCP$_I$ in PhI$_{MAIN}$ and LCI networks from bootstrap subsampling ($n = 1{,}000$) of 100 interactions (two-sided Welch two sample *t*-test). Boxes show the interquartile range (IQR) and median; whiskers show highest and lowest data points within 1.5 IQR; outliers are plotted individually.

differences[11] and screening parameters[12] also affect overall sensitivity. We implemented an interactome mapping framework[6,12] to compare PhI to literature-based network maps from IntAct and BioGRID[13] (LCI$_{BioG}$). Sampling sensitivity of PhI$_{MAIN}$ after five repeat screens was 86 ± 5% (Fig. 1d). For benchmarking, we recurated[12] a positive and a random reference set (PRS$_{PhI}$ and RRS$_{PhI}$) of 92 and 95 protein pairs (Supplementary Table 2), respectively. Benchmarking our Y2H system yielded an unconditional assay sensitivity of 20.4% (Fig. 1e); excluding hormone-dependent PRS$_{PhI}$ interactions increased this to 23%. The resulting overall completion of 16.0 ± 6.8% matches the overlap with LCI data sets (Fig. 1g). Thus, missed interactions explain the incomplete overlap between PhI$_{MAIN}$ and LCI$_{PhI}$, suggesting a low false-discovery rate. This is substantiated by the observation that no RRS$_{PhI}$ pair scored positive (Fig. 1e).

To further assess PhI quality, we used a pulldown assay in which protein pairs are expressed in wheatgerm lysate and, following an anti-Flag immunoprecipitation, interactions are detected through the activity of a second protein fused to Renilla luciferase. Benchmarking this

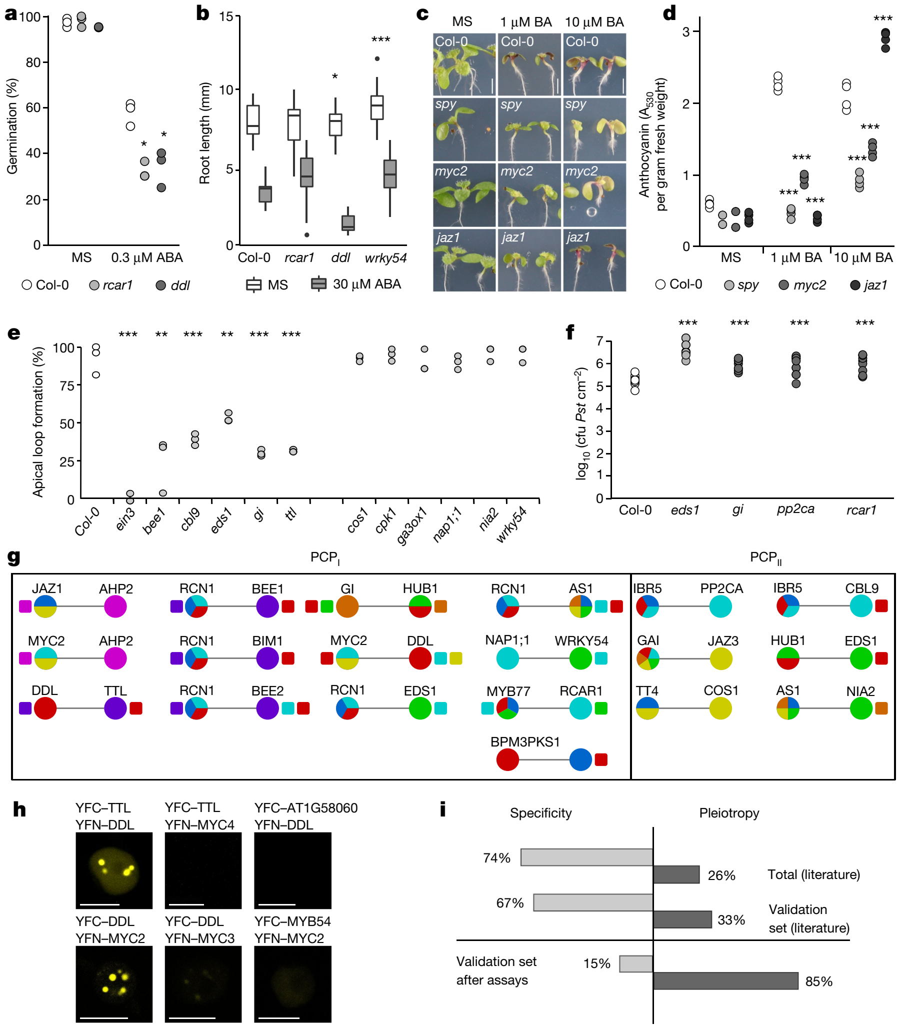


Fig. 2 | Validation of pathway contact points. a, Proportion of germinating seeds in the absence (Murashige and Skoog (MS) standard medium) or presence of 0.3 µM ABA ($n \geq 20$; three repeats). **b**, Root elongation in the absence or presence of 30 µM ABA. Boxes represent the IQR; bold black line represents the median; whiskers indicate highest and lowest data points within 1.5 IQR; outliers are plotted individually ($n \geq 8$; two repeats). Col-0, wild-type background. **c**, **d**, CK-dependent anthocyanin accumulation in the indicated wild-type or mutant backgrounds in response to the indicated concentrations of 6-benzylamino purine (BA). **c**, Seedlings at 10 days after stratification following the indicated treatment. Scale bars, 2 mm. **d**, Quantified anthocyanin content per gram of fresh weight for lines in **c** ($n = 15$; four repeats). A_{530}, absorbance at 530 nm. **e**, ET-induced apical loop formation in response to 10 µM ACC ($n \geq 10$; three repeats). **f**, SA-associated phenotypes in response to inoculation with *Pst*, showing in planta *Pst* titres ($n = 9$); cfu, colony-forming units. **g**, Summary of hormone validation assays for 19 PCPs. Node colours indicate known pathway annotations (Fig. 1a); square colours indicate new phenotypes. **h**, BiFC assay of the PCP$_I$ candidate protein pairs DDL–TTL and DDL–MYC2 and matched negative controls (to the right) (each assay performed in duplicate). Scale bars, 10 µm. **i**, Literature-reported specificity (single pathway annotation) and pleiotropy (multiple pathway annotations) of genes encoding 1,252 total target proteins and 27 proteins in the validation set, as well as updated specificity and pleiotropy following our hormone validation assay. **a**, **b**, **d**–**f**, Two-sided *t*-test; $^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$. **a**–**f**, Precise *P* values, biological repeats and *n* values are shown in Supplementary Table 5.

assay with PRS$_{\mathrm{PhI}}$/RRS$_{\mathrm{PhI}}$ revealed an assay performance similar to that of previous implementations[10,11]; the slightly increased background probably results from the functionally relative coherent search space from which RRS was sampled. Subsequent testing of 285 interactions from the unconditional PhI$_{\mathrm{MAIN}}$, PhI$_{\mathrm{EXT}}$ and PhI$_{\mathrm{Rep\text{-}TF}}$ subsets yielded a PhI validation rate of 22.5%, which is indistinguishable from that of PRS$_{\mathrm{PhI}}$ (23.5%, Fig. 1f) and similar to that for the individual subsets (Extended Data Fig. 1). These data show that PhI is a high-quality map of the *Arabidopsis* phytohormone signalling network and is on a par with high-quality literature data.

For hypothesis generation and local network analyses, the full PhI will be most useful. For topological and systems-level questions, the symmetrically mapped PhI$_{\mathrm{MAIN}}$ should be used to avoid biases[6]. PhI$_{\mathrm{MAIN}}$ has a scale-free degree distribution and—in contrast to LCI$_{\mathrm{PhI}}$ networks—a hierarchical modularity, as expected for unbiased network maps[14] (Fig. 1h and Extended Data Fig. 1). We used PhI$_{\mathrm{MAIN}}$ to investigate the topological organization of phytohormone signalling pathways.

Important features of hierarchical networks are highly connected hubs and interconnected communities[14]. Using a detection algorithm based on edge betweenness[15], we identified 21 network communities in PhI$_{\mathrm{MAIN}}$, of which 9 were substantially enriched in different phytohormone pathways (Fig. 1i, Extended Data Fig. 2 and Supplementary Tables 3, 4). Thus, the topology of PhI$_{\mathrm{MAIN}}$ recapitulates biological knowledge and confirms that at least some pathway proteins are highly interconnected. In addition, most communities encompass proteins from different pathways that possibly mediate crosstalk. In the jasmonate-signalling community, for example, the canonical jasmonate-pathway transcription factor MYC2 is physically linked to ABA signalling via interaction with the protein kinase CIPK14 (Fig. 1j), validated by in vitro pulldown and bimolecular fluorescence

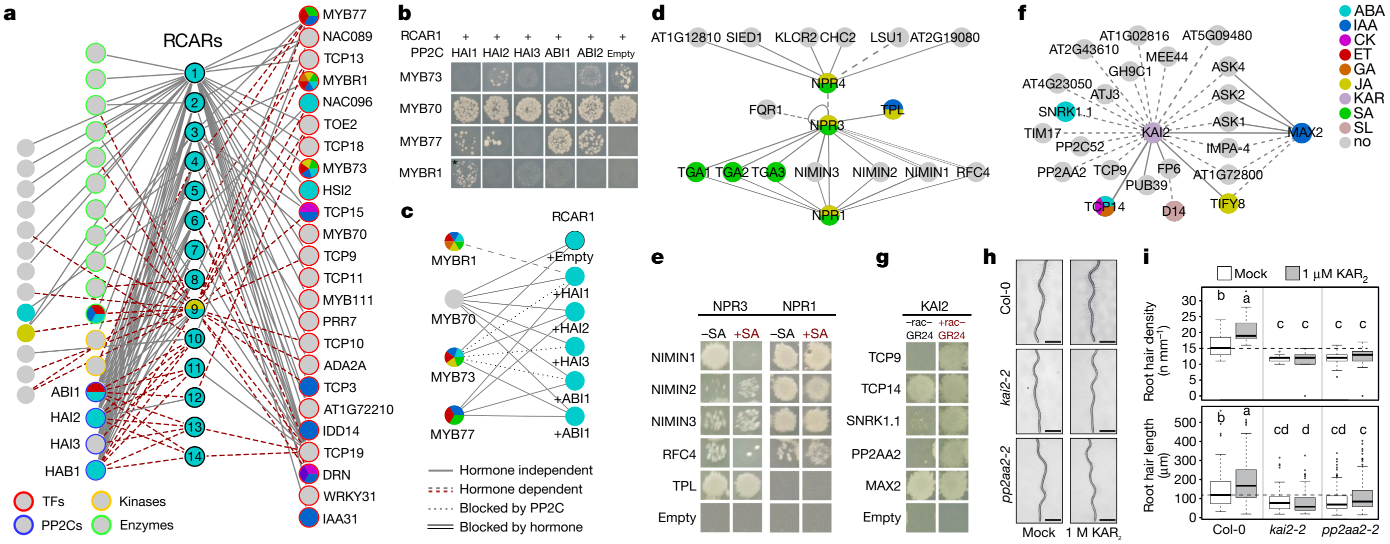


Fig. 3 | Hormone–receptor interactions. a, ABA-dependent Y2H interactions. All identified interactors were systematically tested against all receptors in the presence and absence of ABA. **b**, **c**, Y3H assays for the indicated protein triplets. In all sets, we tested DB–RCAR1 for interactions with AD–MYB proteins in the presence of the indicated PP2Cs and in the presence or absence of ABA. **b**, One of four representative Y3H results. The asterisk indicates an ABA-dependent interaction. **c**, Y3H subnetwork showing the data in **b**. **d**, SA-dependent interactors of NPR1,3,4. **e**, Each image shows one representative yeast colony (from four repeats) in the presence and absence of 100 μM SA for identified NPR interactors. **f**, Hormone-dependent and -independent interactions of KAI2, D14 and MAX2. **g**, Each image shows one representative of four yeast spots for selected KAI2 interactors in presence and absence of rac–GR24. **h**, Representative images from the analysis in **i**, showing root hair phenotypes from the indicated genotypes. Scale bars, 1 mm. **i**, Quantification of root hair density and root hair length after the indicated treatments. Different letters indicate statistically significantly different groups within each graph (one-way analysis of variance (ANOVA); post hoc Tukey, $P \leq 0.05$). Boxes represent IQRs; bold lines show medians; whiskers indicate the highest and lowest data points within 1.5 IQR; outliers are plotted individually. Precise *n* and *P* values for all group comparisons are in Supplementary Table 5. **a**, **c**, **d**, **g**, Modulated interactions are represented according to the key in **c**. Node colours represent hormone annotations as in **g**.

complementation (BiFC) (Extended Data Fig. 2). Additional pathway contacts occur between different communities (Fig. 1j). However, on average only 27% of pathway proteins reside within the corresponding communities, indicating that phytohormone signalling may not be predominantly organized into topological communities (Supplementary Table 3).

We next analysed interpathway connectivity. The distances between the phytohormone pathways are considerably shorter in $\mathrm{PhI}_{\mathrm{MAIN}}$ than in $\mathrm{LCI}_{\mathrm{PhI}}$ (Fig. 1k, l). This is mirrored by the many more pathway contact points (PCPs; protein-interaction-mediated contacts between different pathways) in $\mathrm{PhI}_{\mathrm{MAIN}}$ than $\mathrm{LCI}_{\mathrm{PhI}}$. As some proteins operate in several pathways, we distinguished 192 type I PCPs ($\mathrm{PCP}_{\mathrm{I}}$), involving proteins with strictly different annotations, from 248 type II PCPs ($\mathrm{PCP}_{\mathrm{II}}$), in which the interactors share annotations, but at least one has additional functions (Fig. 1m). Bootstrap subsampling confirmed that $\mathrm{PhI}_{\mathrm{MAIN}}$ contains substantially more type I PCPs (Fig. 1n), but not type II PCPs (data not shown), than do $\mathrm{LCI}_{\mathrm{IntA}}$ or $\mathrm{LCI}_{\mathrm{BioG}}$; this is valid for nearly all pathway pairs (Extended Data Fig. 3). Each discovered PCP supports a specific crosstalk hypothesis, and the abundance of PCPs suggests extensive protein-interaction-mediated information exchange among pathways.

Validation of pathway contact points

We experimentally tested whether PCPs reflect as yet unknown functions of the interacting partners. Assays for most hormones have been well established in seedlings. Therefore, and for standardization, we focused on seedling-expressed PCP interaction pairs. Homozygous T-DNA lines were obtained and validated for 27 candidate proteins involved in 19 PCP pairs. These were used in response assays for seven different phytohormones to establish whether the candidates function in the pathway of their respective PCP partner (Fig. 2a–f, Extended Data Figs. 4–7 and Supplementary Table 5).

ABA regulates seed germination and desiccation stress responses, including root growth[16]. We found that in the presence of 0.3 μM ABA, the germination of wild-type seeds was decreased by roughly 40%. By contrast, the candidate lines *ddl*$_{\mathrm{Lit\_ET}}$ and *eds1*$_{\mathrm{Lit\_SA}}$ displayed a similar ABA hypersensitivity to control seedlings, in which expression of the RCAR1 ABA receptor is disrupted (Fig. 2a and Extended Data Fig. 4). Root growth was significantly altered by 30 μM ABA in 4 additional lines compared with wild type (*as1*$_{\mathrm{Lit\_GA}}$, *bee2*$_{\mathrm{Lit\_BR}}$: $P < 0.01$; *gai1*$_{\mathrm{Lit\_GA}}$, *wrky54*$_{\mathrm{Lit\_SA}}$: $P < 0.001$) (Fig. 2b and Extended Data Fig. 4). Thus, six of the nine candidate lines (66%) exhibited at least one ABA phenotype.

Anthocyanin production is a widely used assay for cytokinin signalling[17]. We found that, at low concentrations, cytokinin-induced anthocyanin accumulation was impaired in the candidate lines to a similar degree as in the *spy* control. At higher cytokinin concentrations, anthocyanin accumulation in the *myc2*$_{\mathrm{Lit\_JA/ABA}}$ line remained similar to that in *spy* seedlings, whereas the *jaz1*$_{\mathrm{Lit\_JA/ABA}}$ line overaccumulated anthocyanin, indicating complexity in cytokinin signalling (Fig. 2c, d).

For the ethylene-based assay, we assayed the ‘triple response’—that is, the formation of exaggerated apical hooks (loops) and the development of shorter and thicker roots and hypocotyls in dark-grown seedlings[18]. Ten of our twelve candidates (83%) displayed an apical loop phenotype; seven of these also displayed a root-growth phenotype, and the *ttl*$_{\mathrm{Lit\_BR}}$ line also had a hypocotyl growth defect following treatment with the ethylene precursor aminocyclopropane carboxylate (ACC) (Fig. 2e and Extended Data Fig. 5). To ensure specificity, we tested seven mutant lines for proteins in PhI that showed no interaction with ethylene-annotated proteins. Of these controls, only one showed a weak root-growth phenotype and none exhibited a hypocotyl or loop formation defect (Fig. 2e and Extended Data Fig. 6).

Salicylic acid mediates defence responses to (hemi-)biotrophic pathogens[19]. Following inoculation with *Pseudomonas syringae* pv. tomato

(*Pst*), titres in the *gi*$_{\mathrm{Lit\_GA}}$ mutant were significantly elevated (Fig. 2f), indicating enhanced disease susceptibility and impaired salicylic acid signalling. Similarly, leaves of mature *rcar1*$_{\mathrm{Lit\_ABA}}$ and *pp2ca*$_{\mathrm{Lit\_ABA}}$ plants supported enhanced *Pst* growth (Fig. 2f). Assays for root-growth inhibition by brassinosteroids, gibberellins and jasmonates revealed new phenotypes for two candidates (for brassinosteroids and gibberellins) or one candidate (for jasmonates) (Extended Data Fig. 4).

Altogether, interactome-guided phenotyping revealed a function in new pathways for 74% of tested proteins (20/27), which are involved in 84% of interactions in the validation set (Fig. 2g and Extended Data Fig. 7). Notably, for all PCP$_{\mathrm{I}}$ pairs a novel function was revealed for at least one partner, such that all interactions are substantiated by phenotypes in at least one common pathway (Fig. 2g). For three of the six PCP$_{\mathrm{II}}$ pairs, an additional common pathway was identified, such that more than half (11/19) of all PCP pairs operate genetically in two common pathways (Fig. 2g). To support these functional data, we used BiFC to demonstrate in planta interactions for nine PCP pairs (Fig. 2h and Extended Data Fig. 7). Before these experiments, a large majority of signalling proteins in the literature and in our validation set were considered to be pathway-specific (Fig. 2i). Following our interactome-guided phenotyping, however, 82% of proteins in the validation set are known to function in multiple pathways, and only one fifth are specific to a single pathway (Fig. 2g, i). The new annotations are distributed across different pathways (Extended Data Fig. 7) and the network degree does not correlate with the number of phenotypes (data not shown). As the validation set is not obviously biased, our observation of widespread pleiotropy may extrapolate to most of the phytohormone signalling network. Thus, our data point to a highly integrated central signal-processing network that channels different inputs into a balanced multifactorial output. To facilitate further studies, we provide an expression-based 'edge score' that indicates the possibility of each PhI interaction occurring in different plant tissues (Supplementary Table 6).

Hormone–receptor interactions

Input into the central processing unit is provided by hormone receptors, which often initiate signalling through small-molecule-regulated protein interactions[20]. To better understand initial phytohormone signalling, we conducted interaction screens using soluble hormone receptors in the presence and absence of their cognate hormone. For the ABA, gibberellic acid, indole-3-acetic acid (IAA), karrikin (KAR), salicylic acid and strigolactone receptors, we identified 241 interactions (PhI$_{\mathrm{HORM}}$), of which 101 are hormone-dependent. Re-identified pairs include interactions of gibberellic acid receptors with DELLA proteins, and of the ABA receptors RCAR/PYR/PYL with type 2C protein phosphatases (PP2Cs) (Fig. 3a and Extended Data Fig. 8), which display known patterns of hormone dependence[21]. Notably, several ABA receptors also interacted with transcription factors and other non-PP2C proteins (Fig. 3a). As some of these also link to PP2Cs, we wondered whether interactions are combinatorially modulated, and investigated by yeast three-hybrid the effect of different PP2Cs on RCAR1/PYL9 interactions with MYB-family transcription factors. The RCAR1–MYB73 interaction was blocked by several PP2Cs, whereas the RCAR1–MYB77 interaction was enabled by ABI1/2, together demonstrating dynamic modulation of complex formation (Fig. 3b, c). In addition, PP2C-independent RCAR functions have been described for RCAR9/PYL6 via MYC2 (ref. [22]) and for RCAR3/PYL8 via MYB77 (ref. [23]). Our data suggest that such core-pathway-independent functions may be more widespread. The independently validated interaction of DELAY-OF-GERMINATION 1 (DOG1) with PP2Cs[24] similarly points to noncanonical PP2C-mediated signalling mechanisms. Thus, core-pathway-independent signalling and complex multimeric interaction regulation are important mechanisms underlying the functional diversification in the ABA signalling system.

Receptors for the defence hormone salicylic acid are the NON-EXPRESSOR OF PATHOGEN RELATED PROTEIN 1 (NPR1) and its orthologues NPR3 and NPR4 (ref. [25]). While NPR1 is a well-studied positive regulator of defence-gene transcription, NPR3 and NPR4 are emerging as alternative negative or complementary transcriptional regulators[25,26]. The pattern of salicylic-acid-regulated NPR3 interactions (Fig. 3d and Extended Data Fig. 9), especially with NIMIN proteins, differs from the described NPR1 pattern[27], suggesting dynamic complexity of this signalling system. EMB1968/RFC4, a member of the replication factor C (RFC) complex, is a new interactor that is common to NPR1 and NPR3, and may integrate defence with DNA repair or replication. Most previously unknown NPR3/NPR4 interactors can be linked to immunity via mutant phenotypes or known interactions with virulence effectors and immune receptors[8] (Fig. 3d and Extended Data Fig. 9). These data support the biological validity of the interactions and indicate that salicylic acid receptors also act through nontranscriptional signalling.

The KAR and strigolactone pathways have been discovered most recently and mediate germination (in the case of KAR) and diverse aspects of development and organismal interactions[28]. We screened the KAR receptor KAI2 and the strigolactone receptor D14, together with the F-box protein MAX2, in the absence and presence of a stereoisomeric mix of two synthetic strigolactones, which bind to D14 and KAI2, respectively[29]. For KAI2 we found the previously described interaction with MAX2 and 21 new interactors, of which 15 are hormone-dependent (Fig. 3f, g and Extended Data Fig. 9). It was previously found that KAI2 regulates root hair length (RHL) and root hair density (RHD)[30]. As both phenotypes are also regulated by auxin, and as the hormone-dependent KAI2 interactor PP2AA2 regulates the PIN auxin exporters, we wondered whether PP2AA2 mediates the effect of KARs on these phenotypes. Similar to *kai2-2* mutants, *pp2aa2-2* plants displayed a lower RHL and RHD than wild-type Col-0 plants (Fig. 3h, i and Supplementary Table 5). Notably, in both *kai2-2* and *pp2aa2-2* mutants the response to exogenous KAR treatment was abolished, indicating that these proteins jointly mediate signalling by the KAR pathway.

Transcriptional changes are common outcomes of phytohormone signalling. Investigating PhI$_{\mathrm{Rep\text{-}TF}}$, we found no evidence of substantial hormone crosstalk at the level of transcriptional regulators from different pathways converging on transcription factors (data not shown). Nonetheless, only a quarter of the transcription factors that we found here to interact with the selected regulators were previously implicated in hormone signalling (Extended Data Fig. 10). Although most pathways converge on TCP-family transcription factors, which are known for their high connectivity[10], most transcription factors interact with repressors from one to three pathways, suggesting more specific signal integration at this level.

We have presented here a systematic map of the *Arabidopsis* phytohormone signalling network, which reveals an unexpectedly high interconnectivity of signalling pathways. If the observed level of functional pleiotropy extends into the larger hormone signalling network, the concept of dedicated signal-transduction pathways may need to be revised in favour of network-based models. The small-molecule-dependent interactions of hormone receptors point towards prominent roles for noncanonical signalling mechanisms. We anticipate that our findings and the PhI resource will stimulate mechanistic and systems-level analyses of *Arabidopsis* and crop plants.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2460-0

1. Krouk, G. et al. A framework integrating plant growth with hormones and nutrients. *Trends Plant Sci*. **16**, 178–182 (2011).
2. Peleg, Z. & Blumwald, E. Hormone balance and abiotic stress tolerance in crop plants. *Curr. Opin. Plant Biol*. **14**, 290–295 (2011).
3. Lumba, S. et al. A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in *Arabidopsis*. *Dev. Cell* **29**, 360–372 (2014).
4. Nemhauser, J. L., Hong, F. & Chory, J. Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell* **126**, 467–475 (2006).
5. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*. **42**, D358–D363 (2014).
6. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
7. Altmann, M., Altmann, S., Falter, C. & Falter-Braun, P. High-quality yeast-2-hybrid interaction network mapping. *Curr. Protoc. Plant Biol*. **3**, e20067 (2018).
8. Weßling, R. et al. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* **16**, 364–375 (2014).
9. Pruneda-Paz, J. L. et al. A genome-scale resource for the functional characterization of *Arabidopsis* transcription factors. *Cell Rep*. **8**, 622–632 (2014).
10. Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333**, 601–607 (2011).
11. Braun, P. et al. An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2009).
12. Braun, P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. *Proteomics* **12**, 1499–1518 (2012).
13. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. **47** (D1), D529–D541 (2019).
14. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*. **12**, 56–68 (2011).
15. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
16. Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R. & Abrams, S. R. Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol*. **61**, 651–679 (2010).
17. Deikman, J. & Hammer, P. E. Induction of anthocyanin accumulation by cytokinins in *Arabidopsis thaliana*. *Plant Physiol*. **108**, 47–57 (1995).
18. Guzmán, P. & Ecker, J. R. Exploiting the triple response of *Arabidopsis* to identify ethylene-related mutants. *Plant Cell* **2**, 513–523 (1990).
19. Vlot, A. C., Dempsey, D. A. & Klessig, D. F. Salicylic acid, a multifaceted hormone to combat disease. *Annu. Rev. Phytopathol*. **47**, 177–206 (2009).
20. Lumba, S., Cutler, S. & McCourt, P. Plant nuclear hormone receptors: a role for small molecules in protein–protein interactions. *Annu. Rev. Plant Biol*. **26**, 445–469 (2010).
21. Tischer, S. V. et al. Combinatorial interaction network of abscisic acid receptors and coreceptors from *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **114**, 10280–10285 (2017).
22. Aleman, F. et al. An ABA-increased interaction of the PYL6 ABA receptor with MYC2 transcription factor: a putative link of ABA and JA signaling. *Sci. Rep*. **6**, 28941 (2016).
23. Zhao, Y. et al. The ABA receptor PYL8 promotes lateral root growth by enhancing MYB77-dependent transcription of auxin-responsive genes. *Sci. Signal*. **7**, ra53 (2014).
24. Née, G. et al. DELAY OF GERMINATION1 requires PP2C phosphatases of the ABA signalling pathway to control seed dormancy. *Nat. Commun*. **8**, 72 (2017).
25. Kuai, X., MacLeod, B. J. & Després, C. Integrating data on the *Arabidopsis* NPR1/NPR3/NPR4 salicylic acid receptors; a differentiating argument. *Front. Plant Sci*. **6**, 235 (2015).
26. Ding, Y. et al. Opposite roles of salicylic acid receptors NPR1 and NPR3/NPR4 in transcriptional regulation of plant immunity. *Cell* **173**, 1454–1467 (2018).
27. Hermann, M. et al. The *Arabidopsis* NIMIN proteins affect NPR1 differentially. *Front Plant Sci* **4**, 88 (2013).
28. Waters, M. T., Gutjahr, C., Bennett, T. & Nelson, D. C. Strigolactone signaling and evolution. *Annu. Rev. Plant Biol*. **68**, 291–322 (2017).
29. Scaffidi, A. et al. Strigolactone hormones and their stereoisomers signal through two related receptor proteins to induce different physiological responses in *Arabidopsis*. *Plant Physiol*. **165**, 1221–1232 (2014).
30. Villaécija-Aguilar, J. A. et al. SMAX1/SMXL2 regulate root and root hair development downstream of KAI2-mediated signalling in *Arabidopsis*. *PLoS Genet*. **15**, e1008327 (2019).

Methods

No statistical methods were used to predetermine sample size. Experiments were not randomized unless otherwise indicated. The investigators were not blinded to allocation during plant phenotyping experiments and outcome assessment.

PhyHormORFeome selection and cloning

We selected target genes as follows: first, those with a known mutant phytohormone signalling phenotype on the basis of annotations in the Arabidopsis Hormone Database (AHD2.0)[31]; second, all members of gene families that were overrepresented in step 1; and third, those highlighted by input from colleagues. In total we selected 1,252 genes, for which 1,226 full-length ORFs could be obtained. To physically assemble the PhyHormORFeome, we picked 688 ORFs from our published AtORFeome collection[8], 276 from the Arabidopsis Biological Resource Center (ABRC), 11 from colleagues and 277 amplified from a Col-0 complementary DNA mix from different tissues. For RNA extraction, we grew 6–10-day-old *A. thaliana* Col-0 seedlings, separating organs and tissues from mature plants (from flowers and siliques (all developmental stages), nodes, internodes, rosette leaves and cauline leaves; roots were from 15-day-old plants; all plants were grown on vertically standing solid MS agar plates imbibed seeds). From all plant organs, tissue types and seedlings, we extracted specific total RNA using a NucleoSpin RNA kit from Macherey and Nagel, following the manufacturer's recommendations. For cDNA synthesis, we modified the Superscript III (Thermo Fisher 18080044) protocol using 25 ng of random primers and 250 ng oligo d(T)16 per 1 µg total RNA. The mixture was heated to 70 °C for 5 min and incubated at 21 °C for 10 min. A mixture of 2.5 µl (0.1 µM) dithiothreitol (DTT), 10 U RNase OUT (40 U µl^{-1}), 250 U SSIII (200 U µl^{-1}), 4 µl SSIII ×5 buffer and 2.5 µl (2 µM) dNTPs was added and incubated at 21 °C for 10 min and then at 42 °C for 120 min. To generate cDNA longer than 5 kilobases, we added a further 250 U of SSIII (200 U µl^{-1}) to the mixture, followed by incubation at 55 °C for 30 min for elongation and 70 °C for 15 min for inactivation. All generated cDNAs from different organs, tissues and seedlings were mixed in equal amounts and 2 µl of the undiluted cDNA mixture (roughly 100 ng) was used to amplify the ORFs of interest. ORF amplification was conducted as a nested polymerase chain reaction (PCR) to attach *attB* cloning sites for further Gateway cloning. The primers comprise 18 base pairs specific to *attB* and 12 bp specific to a partial *attB* site (forward *attB* overhang, GCAGGCTCAGGA; reverse *attB* overhang, GAAAGCTGGGTC). All ORFs were generated with a stop codon. In the second PCR, full *attB* sites were added to the ORFs (forward *attB*, GGGACAAGTTTGTACAAAAAAGCAGGCTCAGGAATG; reverse *attB*, GGGGACCACTTTGTACAAGAAAGCTGGGTC). Gateway cloning and yeast transformation were carried out as described[7]. ORFs cloned herein are available from stock centres.

Y2H-based pipeline for interaction mapping

Network mapping was performed as described[7]. In brief, bait ORFs were expressed as genetic fusions to the GAL4 DNA-binding domain (pDEST–DB); prey ORFs were expressed as genetic fusions to the minimal GAL4 activation domain (pDEST–AD). Both constructs were maintained as low-copy centromeric (*cen*) plasmids and expressed from weak *adh2* promoters. Primary screening was carried out by mating individual DB-plasmid-containing haploid yeast strains (Y8930, MATα) with a mini-pool of haploid Y8800 (MATa) AD-plasmid-containing strains. Following a three-day selection on selective plates containing 1 mM 3-amino-1,2,4-triazole to repress background HIS3 activity, positive single colonies were picked and retested on selective media and cycloheximide control plates. Colonies showing specific selective growth were lysed; the respective ORFs were amplified with generic primers that include position-specific barcodes and subsequently identified using the kiloSeq service by seqWell. All primary Y2H screens were performed once, except for the PhI$_{\mathrm{MAIN}}$ screen, which was carried out five times. The receptor screens and the PhI$_{\mathrm{REP}}$ screen were verified systematically: that is, in the final verification all identified interaction candidates were tested against all receptors or repressors/regulators, respectively. The receptor screens were performed in the absence and presence of the respective phytohormones applied to the selective media. For the ABA-receptor screen, we used 30 µM abscisic acid; for the IAA-receptor screen we used 100 µM IAA; for the gibberellic-acid-receptor screen we used 100 µM GA3; and for the salicylic-acid-receptor screen we used 100 µM salicylic acid. Signalling pathways involving the strigolactone receptor D14 and the karrikin receptor KAI2 were screened with 5 µM rac–GR24.

Y3H assay

We genetically fused RCAR1 to the GAL4 DNA-binding domain using pDEST–DB, and the MYB proteins to the minimal GAL4 activation domain using pDEST–AD. To test for modulation of these interactions, we expressed the indicated PP2Cs from the helper plasmid pVTU–DEST, maintained via the URA3 selection marker. All combinations of RCAR1 and PP2Cs were transformed into the haploid yeast strain Y8930 and mated against Y8800 transformed with the AD–MYB constructs. Y3H assays were performed independently four times in the presence or absence of 30 µM ABA on selective plates (synthetic complete media (Sc) lacking tryptophan (W), leucine (L), uracil (U) and histidine (H) (Sc-W-L-U-H)) containing 1 mM 3-amino-1,2,4-triazole to repress background HIS3 reporter activity. Interactions that were verified in three repeats were counted as Y3H interactions.

Protein–protein interaction reference set

Candidate interactions for the positive reference set (PRS) were compiled from protein–protein interactions from IntAct (downloaded August 2014)[5] and BioGRID (version 3.2.115)[32]. At that time, the IntAct data set contained 17,574 interactions and the BioGRID data set contained 21,474 interactions among *A. thaliana* molecules. From both data sets, we removed protein–DNA interactions, interactions derived from papers that reported more than 100 interactions, and non-binary interactions in protein complexes. Subsequently, we filtered both data sets for interactions described in at least two publications or identified in at least two binary interaction detection methods. This resulted in 233 interactions, from which we randomly picked 140 interactions described in 247 publications for recuration. This yielded a selection of 92 highly reliable binary protein–protein interactions, which constitute the PRS$_{\mathrm{PhI}}$. Ten of these 92 interactions were phytohormone-dependent interactions. To assemble the random reference set (RRS$_{\mathrm{PhI}}$), we randomly sampled 95 protein pairs from proteins in our PhyHormORFeome, excluding already described protein–protein interaction pairs.

Parameters of the interaction mapping framework

To assess the quality of the PhI map—that is, false positive and false negative interactions—we implemented the interactome mapping framework as described[33] and estimated the assay sensitivity, sampling sensitivity, precision and completeness.

The completeness of the PhI$_{\mathrm{MAIN}}$ screening space—that is, the proportion of tested protein pairs in comparison to the theoretical number in the full search space—was based on the number of available ORFs in PhyHormORFeome. The initially defined search space comprised 1,252 loci and thus 1,567,504 possible protein pairs. For the screen of PhI$_{\mathrm{MAIN}}$, we tested 1,254 ORFs corresponding to 1,199 gene loci, of which 1,179 were present as AD- and DB-hybrid constructs, 15 as AD-hybrid constructs only, and 5 as DB-hybrid constructs only. Together, AD- and DB-hybrid constructs for 90.2% of locus combinations were tested for interactions; this percentage corresponds to the completeness.

We estimated the sensitivity of our Y2H assay for detecting phytohormone-signalling-related proteins by benchmarking the system using PRS_{PhI}/RRS_{PhI}. Of the 92 tested possible PRS_{PhI} pairs, 19 pairs were detected, whereas no RRS_{PhI} scored positive, thus yielding an assay sensitivity of 20.7 ± 4.2%. Exclusion of the nine PRS_{PhI} pairs that require presence of a phytohormone—none of which was detected by the unconditional Y2H assay—resulted in an unconditional assay sensitivity of 22.8% ± 4.6%.

Sampling sensitivity—that is, saturation of the screen with regard to detectable interactions measured by the assay sensitivity—was estimated as described[10]. In brief, using a modified Michaelis–Menten function in the R package drc (3.0-1), we modelled the increase of identified interactions per additional repeat screen. Using the first three repeats of the PhI_{MAIN} screen to develop the saturation model, we estimated saturation to occur at 616 ± 38 interactions. We then challenged the model with two further repeats of the primary screen. This resulted in a data set of 529 interactions, which matches the model prediction of 519 ± 31 interactions after five repeats.

Overall sensitivity is the product of assay sensitivity and sampling sensitivity. With an assay sensitivity of 20.7 ± 4.2% and sampling sensitivity of 85.9 ± 5.3%, the overall sensitivity is 17.8% ± 6.8% including conditional interactions in PRS_{PhI}. The unconditional overall sensitivity of 19.4 ± 7.0% is the product of the unconditional assay sensitivity of 22.8 ± 4.6% and the sampling sensitivity of 85.9 ± 5.3%. We estimated the overall completion of the screen as the product of the overall sensitivity and completeness; thus, overall completion of PhI_{MAIN} is 16.0% ± 6.8%.

Luciferase validation assay

Protein expression. We expressed proteins constituting PRS_{PhI}/RRS_{PhI} pairs and the interaction pairs from the different subsets (see text) in cell-free coupled transcription translation wheatgerm lysate (Promega, L3260) using SP6 promoters. Of each protein pair, one partner was expressed with an amino-terminal Flag tag, and the second carried an N-terminal Renilla luciferase fusion. Protein pairs were coexpressed according to the manufacturer's protocol (Promega), except that the amounts were proportionally adjusted to a final reaction volume of 20 µl. Input DNA plasmids were isolated from 1 ml bacterial cultures grown in Terrific Broth for 20 h on a vibration platform shaker (Union Scientific) using a Qiagen Biorobot3000 and Turbo Prep 96-well plasmid isolation kits. These yielded approximately 20–40 ng μl^{-1} of DNA, of which 4 µl were used in a 20 µl (final volume) TnT reaction. Protein expression was carried out by incubating the reaction mixture containing both plasmids for 2 h at 30 °C.

Preparation of plates for immunoprecipitation. Plates coated with anti-Flag antibodies were made in-house by incubating white 96-well Lumitrac high binding plates (Greiner) overnight at 4 °C with 75 µl phosphate-buffered saline (PBS; pH 7.4) per well containing 8 µg ml^{-1} M2 anti-Flag antibody (Sigma). Two hours before use, the antibody solution was replaced with 100 µl blocking buffer containing 10 µg μl^{-1} bovine serum albumin (BSA), followed by two hours of shaking at room temperature.

Interaction detection. Following protein expression, 2 µl of lysate were diluted in 28 µl PBS (pH 7.4); expression of the prey protein was quantified by adding 10 µl Renilla glow luciferase substrate. The remaining expression lysate was diluted in 42 µl blocking buffer and added to the empty wells of the immunoprecipitation plates. The plates were incubated with gentle shaking for 2 h at 4 °C, then washed three times with 100 µl blocking buffer. Coimmunoprecipitation efficiency was determined by adding 10 µl Renilla glow luciferase substrate (Promega) diluted in 30 µl PBS (pH 7.4). Interaction pairs were scored as positive when the expression level was at least 10% of the median of the respective plate (expression positive), the immunoprecipitation exceeded the median immunoprecipitation of the plate (minimum immunoprecipitation signal), and the *Z*-test of the immunoprecipitation efficiency gave a score greater than 0.4 (immunoprecipitation ratio of the sample relative to that of the plate). To determine data set precision, we tested a total of 446 pairs from PRS_{PhI} (78), PRS_{unc} (69), RRS_{PhI} (83), PhI_{MAIN} (115), PhI_{EXT} (110) and $PhI_{Rep\text{-}TF}$ (60). Data set differences were statistically compared using a one-sided Fisher's exact test.

Network topology

To determine the network topology of PhI_{MAIN}, we calculated the distributions of degree and clustering coefficients for the indicated networks using the igraph package. We used the distributions to determine the underlying network topology[34].

Network visualization and annotation

Networks were visualized with Cytoscape[35] (version 3.7.2) using protein annotations from Araport11 (ref. [36]). Hormone annotations were downloaded from AHD2.0, and extracted from The Arabidopsis Information Resource (TAIR10) gene ontology annotations (downloaded on 3 August 2018). Hormone annotations were inferred from gene ontology annotations when a gene had a gene ontology term that contained one of these key words: auxin, abscisic acid, brassinosteroid, cytokinin, ethylene, gibberellin, jasmonic acid, salicylic acid, strigolactone or karrikin. Gene ontology annotations with evidence code IEP were excluded from all analyses.

Community detection

Communities in PhI_{MAIN} were determined using the edge-betweenness algorithm[15] implemented in the *R* package igraph (version 1.2.4)[37].

Hormone enrichment

Communities were tested for enrichment with proteins that function in hormone signalling pathways by using the hormone annotations from AHD2.0 and TAIR10. For each community, we compared the number of proteins with a given pathway annotation to the total in the full PhI_{MAIN} network using a two-sided Fisher's exact test and multiple hypothesis corrected with Benjamini–Hochberg algorithm.

Gene ontology enrichment

All communities were tested for gene ontology enrichment using the *R* package GOstats (version 2.50.0)[38]. Gene ontology annotation data were derived from the *R* package GO.db (version 3.7.0). Communities were tested for overrepresentation of gene ontology terms by using a hypergeometric test function, hyperGTest, invoked with parameter conditional = TRUE. *P* values of each community were corrected for testing multiple gene ontology terms using the Benjamini–Hochberg method.

Pathway distance calculation

To determine the distance between different hormone pathways, we determined all shortest paths between proteins of the respective hormone signalling pathways. We considered only those shortest paths that do not contain proteins in the same pathways as those under consideration. We calculated the mean path length from all shortest paths between the two pathways.

PCP determination and network comparison

We used hormone pathway annotations from AHD2.0 and Gene Ontology for this analysis. From the PhI_{MAIN} network, we extracted interactions between two proteins annotated with distinct hormone signalling pathways (type I), and interactions between two proteins involved in distinct but also common pathways (type II). To compare the number of PCPs in PhI_{MAIN} with LCI networks, we used a subsampling bootstrapping approach. From each network we conducted 1,000 iterations of sampling 100 interactions without replacement. For each sampling, we determined the total number of type I and type II PCPs and the number of PCPs for each specific hormone combination. We compared the derived distributions for total PCPs from PhI_{MAIN} with the distributions obtained from LCI networks using a two-sided Welch's two-sample *t*-test. We compared the distributions of hormone-combination-specific PCPs using a two-sided Wilcoxon's test and multiple testing corrected by the number of hormone combinations tested (45).

Literature curated interactions

Interactions curated from literature were downloaded from IntAct[5] and BioGRID[39]. *Arabidopsis* protein–protein interactions were extracted from the IntAct database (downloaded in June 2016) and from the BioGRID database version 3.4.142 (downloaded in November 2016).

Phytohormone sources

Phytohormones were obtained from the following manufacturers: ACC, Sigma (catalogue number A-3903); 6-benzylamino purine, Sigma (B3408); brassinolide, Sigma (B1439); karrikin2 (KAR2), Olchemim (025 682); karrikin2 (KAR2), Toronto Research Chemicals (F864800) (for Y2H experiments); gibberellic acid 3, Duchefa (G0907); rac–GR24, Chiralix (CX23880); IAA, Sigma (I2886); paclobutrazol (Pac), Duchefa (P0922); salicylic acid, Sigma (S5922); ABA, Sigma (A1049); and methyl-jasmonate (Me-JA), Sigma (392707).

Plant material and growth conditions

All *A. thaliana* lines—wild-type, *ahp2*, *as1*, *bee1*, *bee2*, *bim1*, *bpm3*, *cbl9*, *cos1*, *cpk1*, *ddl*, *eds1*, *ga3ox1*, *gai*, *gi*, *hub1*, *ibr5*, *jaz1*, *jaz3*, *kai2-2*, *myb77*, *myc2*, *nap1;1*, *nia2*, *pks1*, *pp2aa2-2*, *pp2ca*, *rcar1*, *rcn1*, *rgl1*, *tt4*, *ttl*, *wrky54*, *rga*, *rga-28*, *spy* and *ein3*—are in the Col genetic background. Seeds were obtained from the Nottingham Arabidopsis Stock Centre (NASC) and propagated for three generations in a greenhouse environment at 21 °C and LD light (16 h/8 h). For genotyping, one leaf of a 12–14-day-old plant was frozen in liquid nitrogen, and genomic DNA was extracted in 1.5 ml tubes using Edwards DNA-extraction buffer[40]. For expression-level analysis of mutant lines, RNA was extracted using a NucleoSpin RNA kit from Macherey–Nagel and the Moloney murine leukemia virus (M-MuLV) reverse transcriptase (Biozym 350400201) according to the manufacturer's recommendations. All seeds were surface sterilized and stratified for 3 days at 4 °C in the dark on MS plates or on plates containing the indicated additives. LD light conditions were 75–85 $\mu M\ m^{-2}\ s^{-1}$ measured with a LI-250A light sensor (LI-COR). *Nicotiana benthamiana* seeds were spread on soil and grown in a greenhouse environment at 23 °C and with LD light (16 h/8 h). For all assays, measurements were carried out with distinct samples (no repeat measurements on the same sample). For statistical tests of significance, we assumed a normal distribution of the measured variable (for example, root length); hormone treatments and genotype were tested as covariates.

Measurement of triple response to ethylene

Sterile seeds were placed directly on standard MS plates or on plates with 10 μM ACC, then stratified for 3 days at 4 °C in the dark, transferred into light for 1 h to induce germination, and incubated for 3 days at 23 °C in the dark. The formation of apical hooks versus loops was scored visually; image analysis for hypocotyl and root-length determination was performed using Fiji imaging software[41] and the Simple Neurite Tracer[42] plugin (version 3.1.3).

Root elongation measurements

Seedlings were grown on MS plates to five days after germination, and then transferred to MS mock plates or MS containing the appropriate phytohormone additive as indicated in the figures (Pac, 0.5 μM, 1.0 μM; brassinolide, 0.1 μM, 0.5 μM; Me-JA, 25 μM). Transferred seedlings were grown in the vertical position for another four days at 23 °C in LD light conditions (16 h/8 h). Root lengths were determined as above.

Anthocyanin accumulation

Anthocyanin content in response to the indicated treatments was determined as described in ref. [43] and expressed per gram of fresh weight.

Root hair growth

Root hair growth was analysed as in ref. [30] using 1 μM KAR2. *Arabidopsis* seeds were stratified in the dark for 3 days at 4 °C and then transferred to a growth cabinet at 22 °C with a 16 h/8 h light/dark cycle (intensity approximately 100 $\mu M\ m^{-2}\ s^{-1}$). Images were taken with a Zeiss SteREO Discovery.V8 microscope (Carl Zeiss) equipped with a Zeiss Axiocam 503 colour camera (Carl Zeiss). We determined the number of root hairs by counting root hairs with lengths of 2–3 mm (from the root tip) on each root, and root hair length was measured for 10–12 different root hairs per root as described above. For karrikin treatments, KAR2 (Olchemim) was dissolved in 75% methanol to prepare a 10 mM stock solution. Analysis and data were based on two repeats.

Infection assay

To measure bacterial proliferation in 4–5-week-old plants, we carried out assays as described[44] using *Pseudomonas syringae* pv. tomato DC3000. To prepare the inoculum, bacteria were grown overnight on NYGA medium (5 $g\ l^{-1}$ bactopeptone, 3 $g\ l^{-1}$ yeast extract and 20 $ml\ l^{-1}$ glycerol) and resuspended and diluted to 1×10^5 colony-forming units per ml in 10 mM $MgCl_2$. Bacteria were inoculated by syringe infiltration of two leaves per plant, and harvested at four days post inoculation as described[44]. In short, three leaf discs per sample were incubated for 1 h in 10 mM $MgCl_2$ containing 0.01% Silwett. The resulting suspension was serially diluted, 20 μl of each dilution were plated, and colonies were counted after two days.

BiFC assay

For BiFC, we used the vectors pMDC43–YFC, pMDC43–YFN[45] and pDEST-VYNE(R), pDEST-VYCE(R)[46]. After Gateway recombination, the ORF-containing destination clones were introduced into *Agrobacterium tumefaciens* strain GV3101. Transformed *A. tumefaciens* cells were grown overnight and resuspended in infiltration buffer (10 mM $MgCl_2$, 10 mM 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 5.6 and 150 μM acetosyringone) with a final optical density at 600 nm (OD_{600}) of 0.3 for each expression vector. The abaxial leaf surface of *N. benthamiana* plants was transiently transformed with *A. tumefaciens* (containing the constructs and the silencing inhibitor protein p19) by infiltration using a needleless syringe. Two days after infiltration, two leaves from two independently transformed plants were used for fluorescence detection. Reconstitution of fluorescence was observed under an epifluorescence microscope (Olympus BX61) using yellow fluorescent protein (YFP) and red fluorescent protein (RFP) bandpass filters for the YFC–MYC2 and YFN–CIPK14 interactions, respectively; either a TCS SP8 multiphoton microscope (Leica) or an LSM880 laser scanning confocal microscope (Carl Zeiss) was used for the remaining BiFC assays. The laser excitation wavelength for both microscopes was 488 nm and the detection band was set to 493–545 nm for Venus protein. The objectives were a PL APO ×40/1.10 and a Plan-Apochromat ×20/0.8 M27 for the TCS SP8 and LSM880, respectively. Image analysis was performed using Fiji imaging software[41]. Analyses were performed in duplicate for all constructs.

In vitro pulldown assays

For in vitro pulldown assays, amylose resin (New England Biolabs) coated with MBP–MYC2 was incubated for 2 h at 4 °C with an equimolar amount of purified glutathione-*S*-reductase (GST)–CIPK14. Wash and elution steps were performed according to the manufacturer's instructions. Pulldowns were analysed by western blot using antibodies against GST (Amersham Biosciences) and maltose-binding protein (MBP, New England Biolabs).

Protein–protein interaction likely scores

We developed an edge-score model to determine the protein–protein interaction likely score in different plant tissues and development states. The edge-score modelling was designed to exploit transcript abundance to estimate the possibility, and to some extent the likelihood, of an interaction taking place in a given tissue and condition. It is based on using transcript abundance as a proxy for protein concentration and modelling binary complex formation by the law of mass action. Tissue-specific transcriptome data were collected from ref. [47]. FastQC (version 0.11.7) was used for read quality control before and after trimming. Adaptor sequences and low-quality reads were trimmed with Trimmomatic version 0.36 (ref. [48]) using the ILLUMINACLIP: TruSeq3-SE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDING WINDOW:4:15 and MINLEN:36 options. High-quality reads were mapped to the TAIR10 reference genome. The estimation of gene abundance was performed with Kallisto version 0.45 (ref. [49]). To estimate the chance of two proteins, i and j, interacting in a given condition, we used the law of mass action to obtain a quantitative estimate of their interaction feasibility. We estimated the amount of proteins i and j using their respective transcript levels, t_i and t_j, as proxies, and determined edge scores as follows. In each tissue, let $t_i^{t_k}$ and $t_j^{t_k}$ denote the abundance of genes i and j in tissue t_k. The score of the interaction between proteins i and j in tissue t_k is calculated as:

$$S_{ij}^{t_k} = t_i^{t_k} \times t_j^{t_k}$$

After obtaining a score for each interaction in each tissue, we compute the raw edge score (es′) of a specific interaction in tissue t_k by Z-transformation:

$$\mathrm{es}'^{t_k}_{ij} = \frac{S_{ij}^{t_k} - \overline{S_{ij}^{t_k}}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(S_{ij}^{t_k} - \overline{S_{ij}^{t_k}}\right)^2}}$$

Finally, we normalize this score to the range of [0, 1]:

$$\mathrm{es}^{t_k}_{ij} = \frac{\mathrm{es}'^{t_k}_{ij} - \min(\mathrm{es}'^{t_k}_{ij})}{\max(\mathrm{es}'^{t_k}_{ij}) - \min(\mathrm{es}'^{t_k}_{ij})}$$

A higher normalized edge score indicates that an interaction in this tissue is more likely, as both proteins are expressed jointly.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All functional, genetic and interaction data generated here are available as Supplementary Information. The genes selected for interactome mapping (the search space) are presented in Supplementary Table 1. All protein–protein interaction data can be found in Supplementary Table 2. The data for genetic validation assays can be found in Supplementary Table 5. The preliminary edge scores for all interactions identified here are in Supplementary Table 6. Additionally, all protein interactions identified here have been submitted to IMEx (http://www.imexconsortium.org) through IntAct[5] with identification code IM-27834. Source data are provided with this paper.

Code availability

Custom scripts used here are available at https://github.com/INET-HMGU/PhyHormInteractome.

31. Jiang, Z. et al. AHD2.0: an update version of *Arabidopsis* Hormone Database for plant systematic studies. *Nucleic Acids Res*. **39**, D1123–D1129 (2011).
32. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. **34**, D535–D539 (2006).
33. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
34. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*. **5**, 101–113 (2004).
35. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. **13**, 2498–2504 (2003).
36. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. **89**, 789–804 (2017).
37. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst*. **1695**, 1–9 (2006).
38. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
39. Chatr-aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res*. **45** (D1), D369–D379 (2017).
40. Edwards, K., Johnstone, C. & Thompson, C. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res*. **19**, 1349 (1991).
41. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
42. Longair, M. H., Baker, D. A. & Armstrong, J. D. Simple Neurite Tracer: open source software for reconstruction, visualization and analysis of neuronal processes. *Bioinformatics* **27**, 2453–2454 (2011).
43. Nakata, M. & Ohme-Takagi, M. Quantification of anthocyanin content. *Bio Protoc*. **4**, e1098 (2014).
44. Wenig, M. et al. Systemic acquired resistance networks amplify airborne defense cues. *Nat. Commun*. **10**, 3813 (2019).
45. Belda-Palazón, B. et al. Aminopropyltransferases involved in polyamine biosynthesis localize preferentially in the nucleus of plant cells. *PLoS ONE* **7**, e46907 (2012).
46. Gehl, C. et al. Quantitative analysis of dynamic protein-protein interactions in planta by a floated-leaf luciferase complementation imaging (FLuCI) assay using binary Gateway vectors. *Plant J*. **67**, 542–553 (2011).
47. Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J*. **88**, 1058–1070 (2016).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*. **34**, 525–527 (2016); erratum **34**, 888 (2016).

Acknowledgements We thank all INET members for helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether program (grant GU1423/1-1) to C.G.; by DFG grant SFB924 to A.C.V., C.G., E.G. and P.F.-B.; by the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry for Education, Science, Research and Technology) grant 031L0141 e:Bio-Modul III: ChlamyInt to P.F.-B.; and by the European Research Council's Horizon 2020 Research and Innovation Programme (grant agreement 648420) to P.F.-B.

Author contributions The authors contributed as follows: project conception, P.F.-B.; ORF selection and cloning, M.A., S.A., P.F.-B., G.W.B., S.C. and C.G.; Y2H screening, M.A., P.A.R., L.E.V., M.S., V.Y., R.P. and N.M.R.; hormone-dependent Y2H screens, M.A. and J.P.; $\mathrm{PRS}_{\mathrm{PhI}}$/$\mathrm{RRS}_{\mathrm{PhI}}$ curation, M.A., S.A., N.M.R., A.G.-M and P.F.-B.; network analyses, S.A., P.F.-B., M.A., K.G.K. and K.F.X.M.; edge-score calculation, C.-W.L., S.A. and P.F.-B; pulldown experiment, J.P. and N.M.R.; systematic validation assay, B.W., P.A.R., M.S., M.A., A.S., V.Y. and P.F.-B.; BiFC validations, Fig. 2h and Extended Data Fig. 7b, P.A.R., M.A., N.M.R. and J.P.; genetic validation assays, for ABA, M.A., E.G., L.G., S.A. and P.F.-B.; for cytokinins, M.A. and L.E.V.; for ethylene, M.A. and L.E.V; for salicylic acid, A.C.V., M.W., J.S. and M.A.; for jasmonate, N.M.R., M.A. and P.A.R.; for gibberellic acid, M.A.; for karrikin/GR24 vignette and Fig. 3h, j, J.A.V.-A. and C.G.; figures, M.A., S.A., C.F. and P.F.-B.; manuscript writing, P.F.-B., C.F., M.A., S.A., C.G. and A.C.V.

Competing interests The authors declare no competing interests.

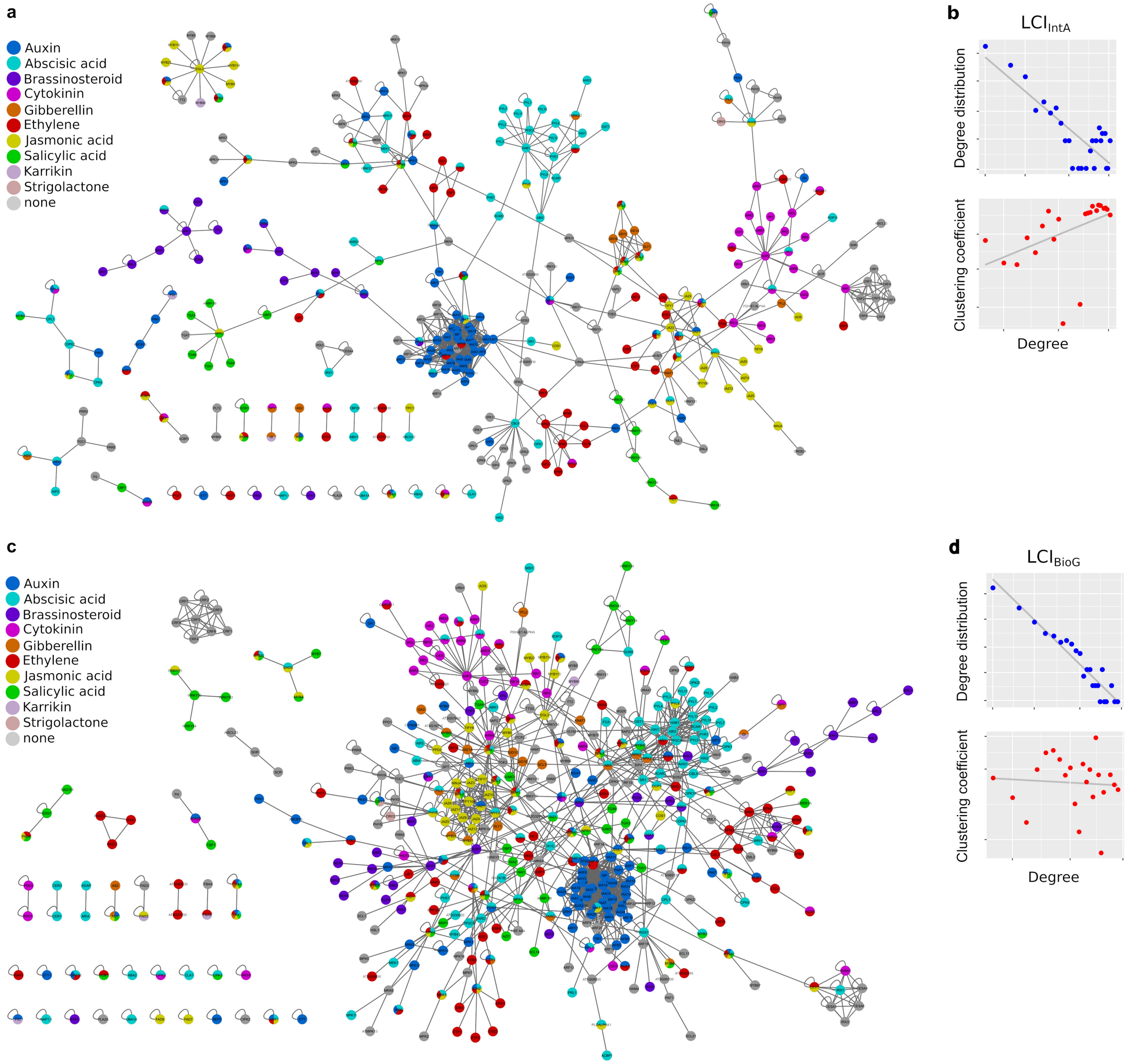
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2460-0.

Correspondence and requests for materials should be addressed to P.F.-B.

Peer review information *Nature* thanks Peter McCourt, Ulrich Stelzl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

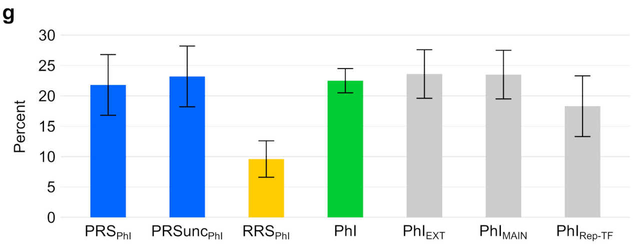


e

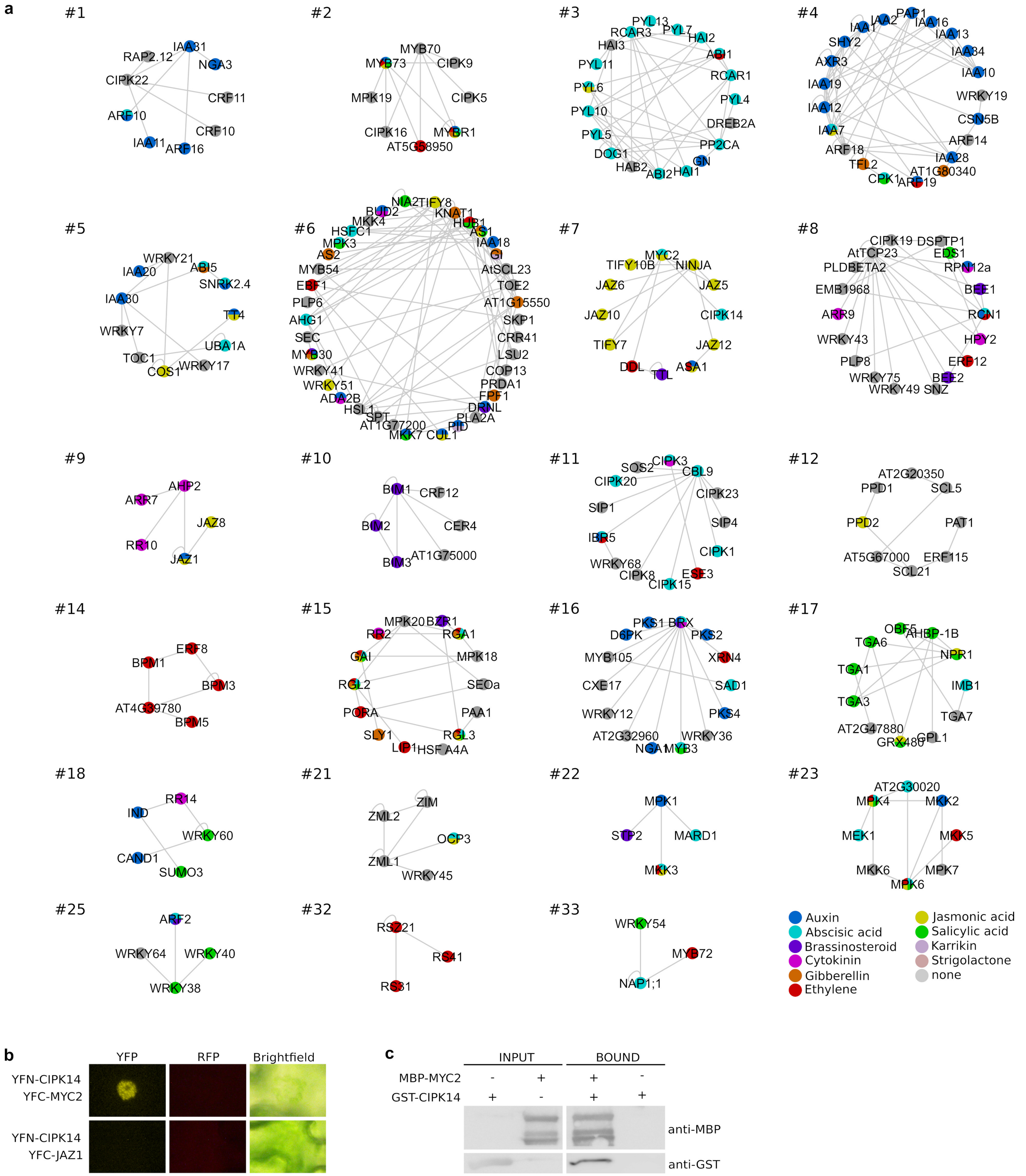
Network	Proteins	Interactions
PhI_{MAIN}	273	529
PhI_{EXT}	658	789
$PhI_{Rep\text{-}TF}$	141	690
PhI_{HORM}	138	241
PhI	924	2072

f

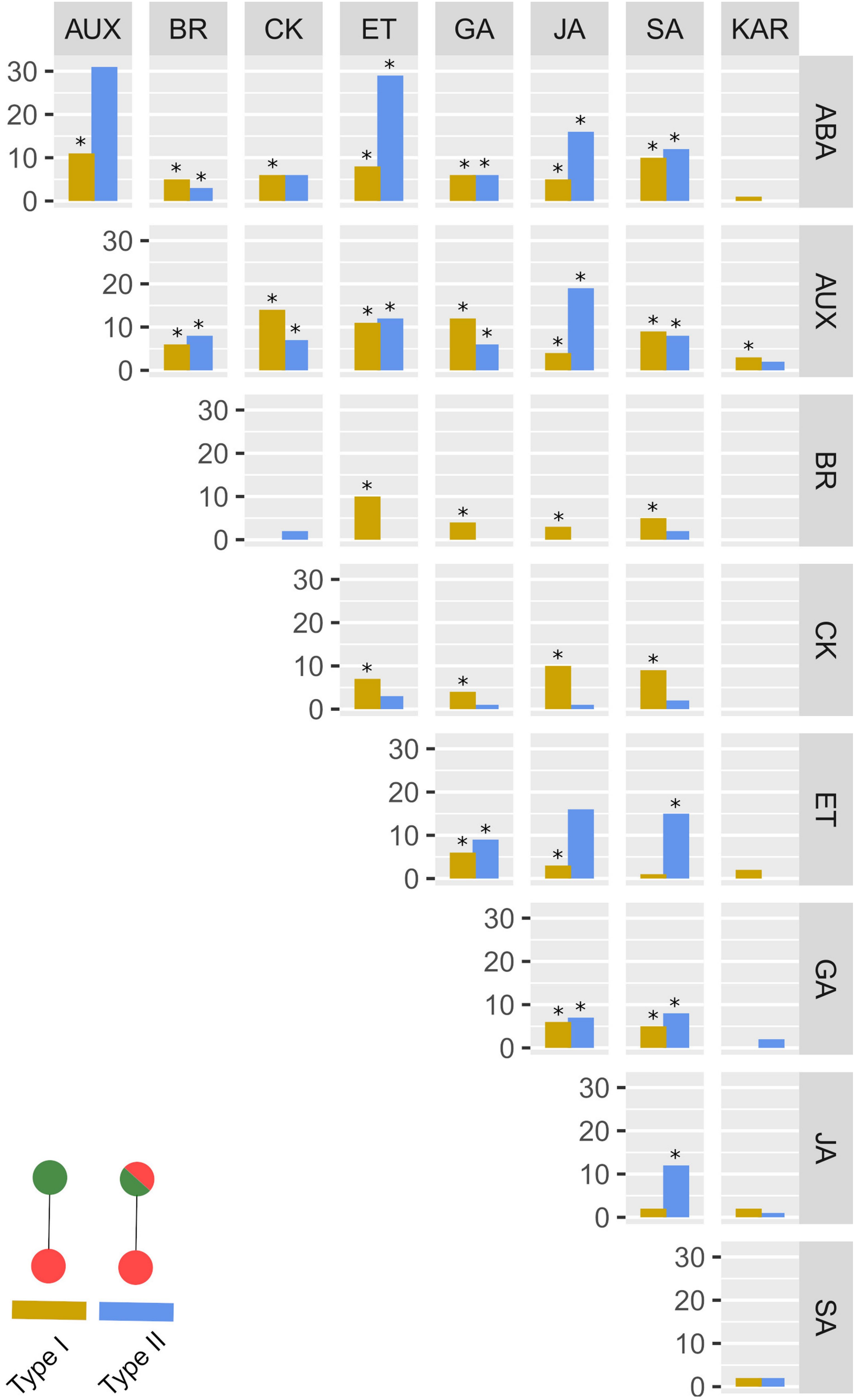
Hormone		Interactions total	 new
ABA	●	731	581
AUX	●	981	705
BR	●	180	155
CK	●	233	204
ET	●	521	435
GA	●	239	201
JA	●	594	475
KAR	●	35	34
SA	●	410	328
SL	●	1	1
PhI		2072	1572



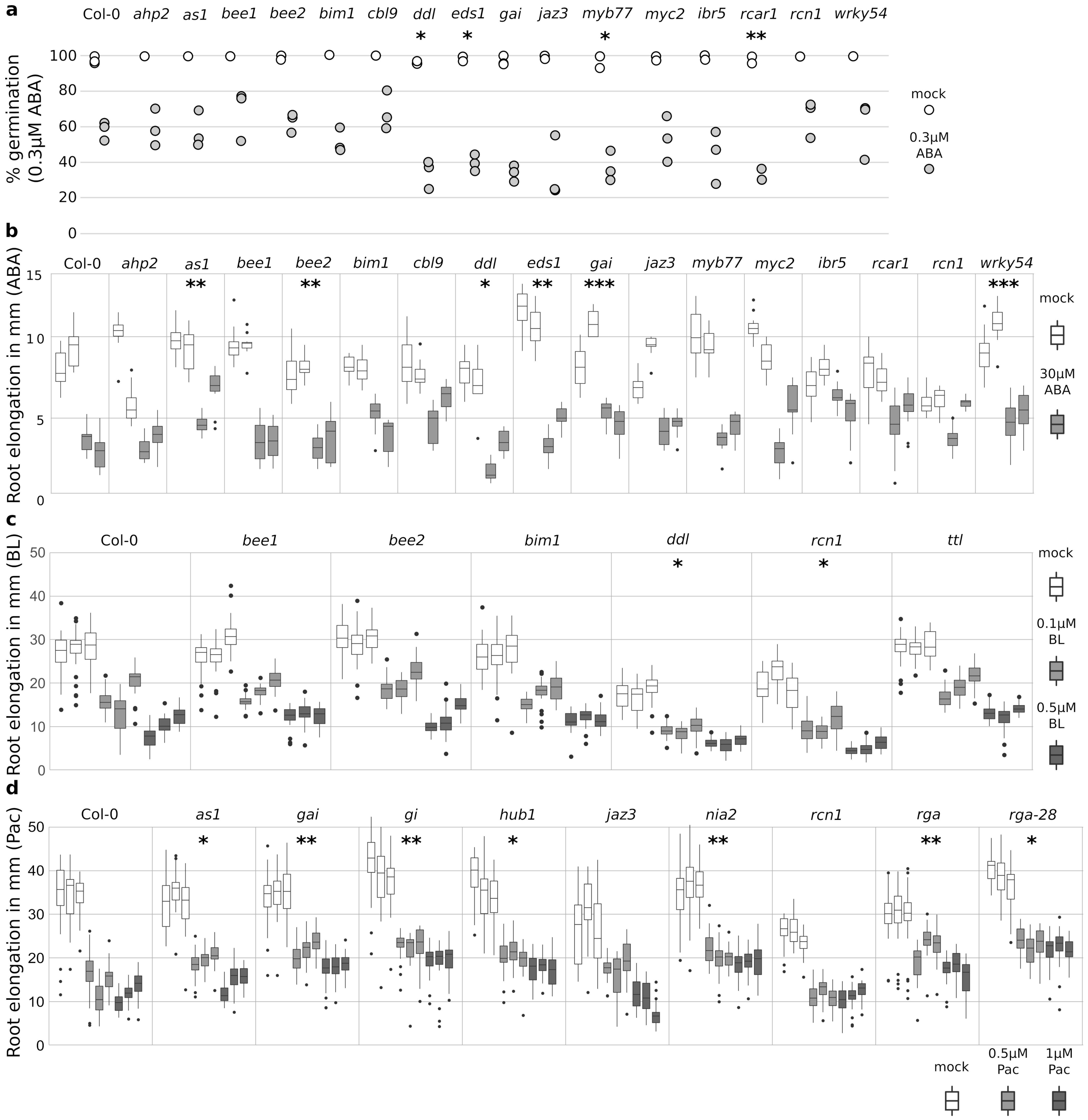
Extended Data Fig. 1 | Network analyses. a, Network map showing binary protein–protein interactions among search space proteins derived from IntAct (LCI_{IntA}). The colour code indicates existing hormone pathway annotations, as indicated in the key. **b**, Degree distribution and clustering coefficient distribution (on a log–log scale) for the network in **a**. **c**, Network map showing binary protein–protein interactions among search space proteins derived from BioGRID (LCI_{BioG}). **d**, As in **b**, but for network shown in **c**. **e**, Number of proteins and interactions in the PhI interactome subsets. **f**, Number of total and new interactions found in PhI for all proteins belonging to each pathway, and the non-redundant total for PhI. **g**, Fraction of positive scoring pairs in PRS_{PhI} (78), PRS_{unc} (hormone-independent PRS interactions) (69), RRS_{PhI} (85), combined PhI subsets (green; 285) and the individual subsets from the single Y2H screens: PhI_{EXT} (110), PhI_{MAIN} (115) and Rep-TF (60). Error bars indicate standard errors of proportions. Individual results for all pairs are provided in Supplementary Table 2.



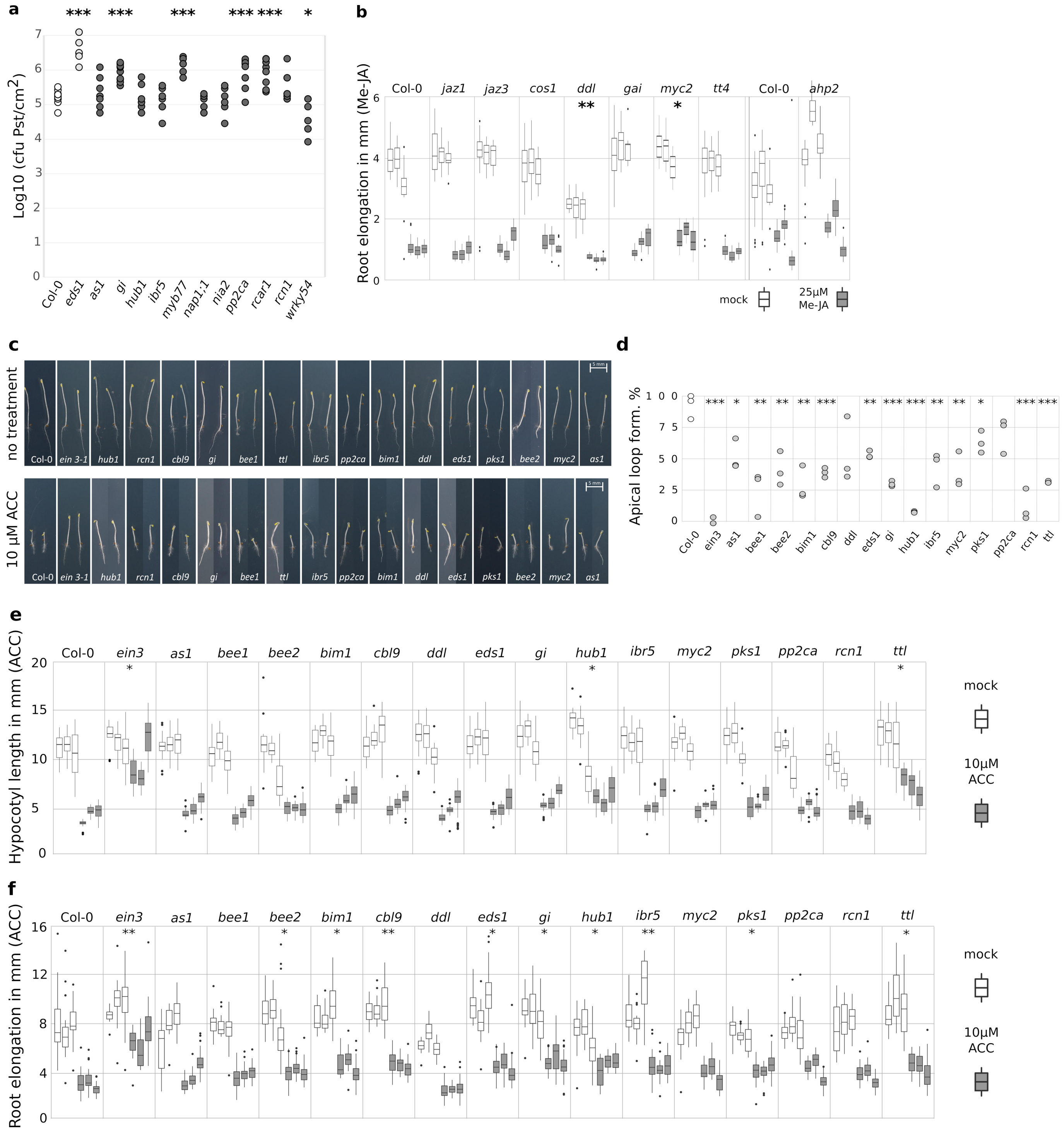
Extended Data Fig. 2 | Communities and validation. a, Communities with three or more nodes identified in $\mathrm{PhI}_{\mathrm{MAIN}}$. Community numbers (1–33) correspond to the numbering in Supplementary Table 3. The colour code indicates existing hormone pathway annotations as indicated in the legend. Node labels are gene symbols when available, and otherwise are locus identification numbers. **b**, BiFC analysis of CIPK14–MYC2. *Nicotiana benthamiana* epidermal leaves that are transiently coexpressing cYFP–MYC2 (cYFP: C-terminal amino-acid residues 154–240 of YFP) and nYFP–CIPK14 (nYFP: N-terminal amino-acid residues 1–154 of YFP) restore YFP fluorescence, whereas coexpression of the noninteracting cYFP–JAZ1 and nYFP–CIPK14 does not. **c**, Maltose-binding-protein (MBP) pulldown of MBP–MYC2 and of CIPK14 tagged with glutathione-*S*-reductase (GST) shows specific copurification of the latter. **b**, **c**, Representative results of two experiments with similar results are shown.



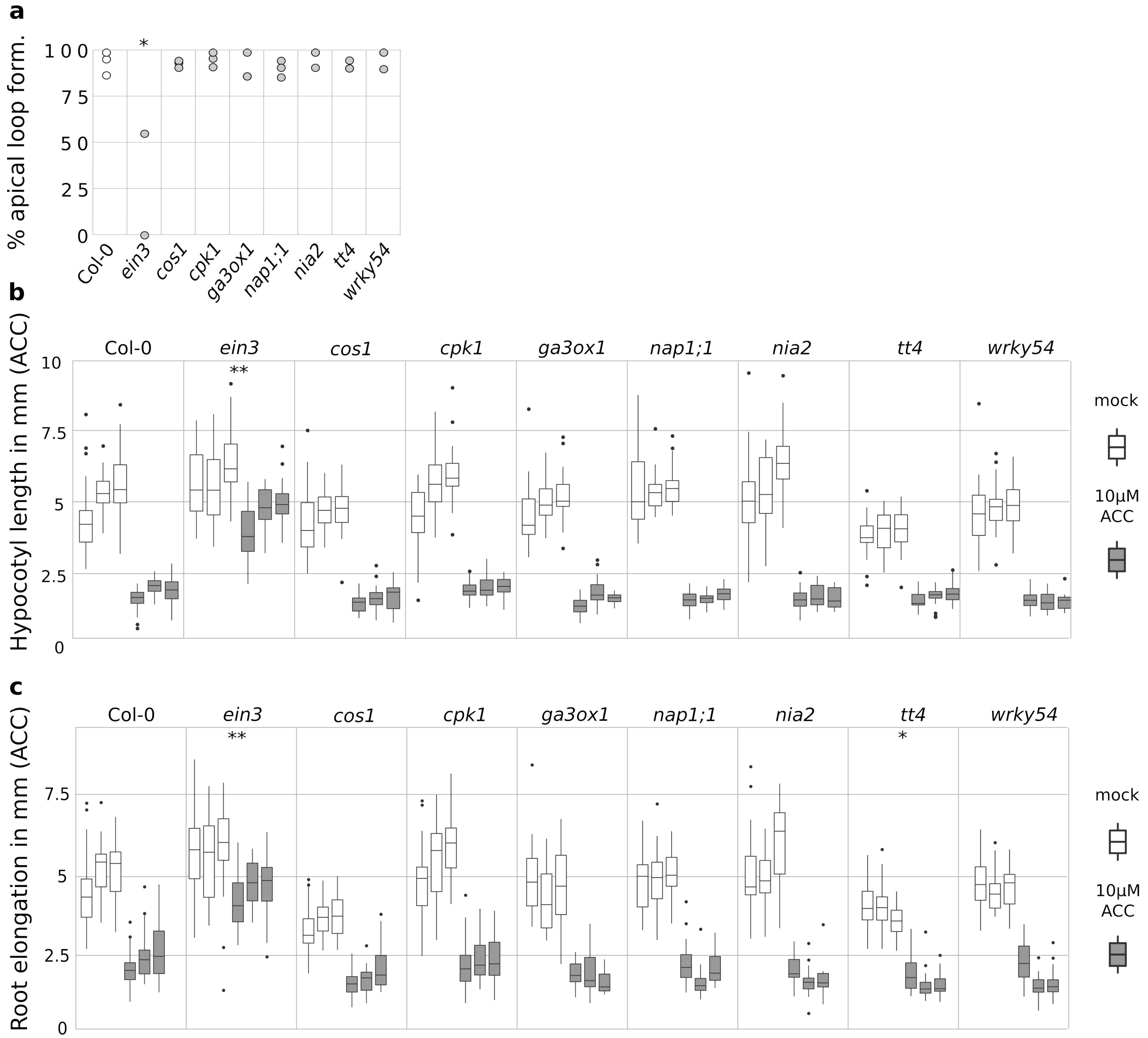
Extended Data Fig. 3 | Enrichment of pathway contact points. Number of PCP_I and PCP_{II} per hormone combination. Asterisks indicate a significantly higher number of PCPs compared with LCI_{IntA} ($*P \leq 0.001$), as obtained by bootstrap subsampling analysis ($n = 1,000$) of 100 interactions followed by two-sided Welch's two-sample *t*-test. Precise *P* values for PCP_I and PCP_{II} and pathway combinations are listed in Supplementary Table 2. The key at the bottom illustrates PCP_I (counts in brown bars) as interactions between proteins with strictly different pathway annotations, and PCP_{II} (counts in blue bars) as interactions of proteins with both shared and different pathway annotations.



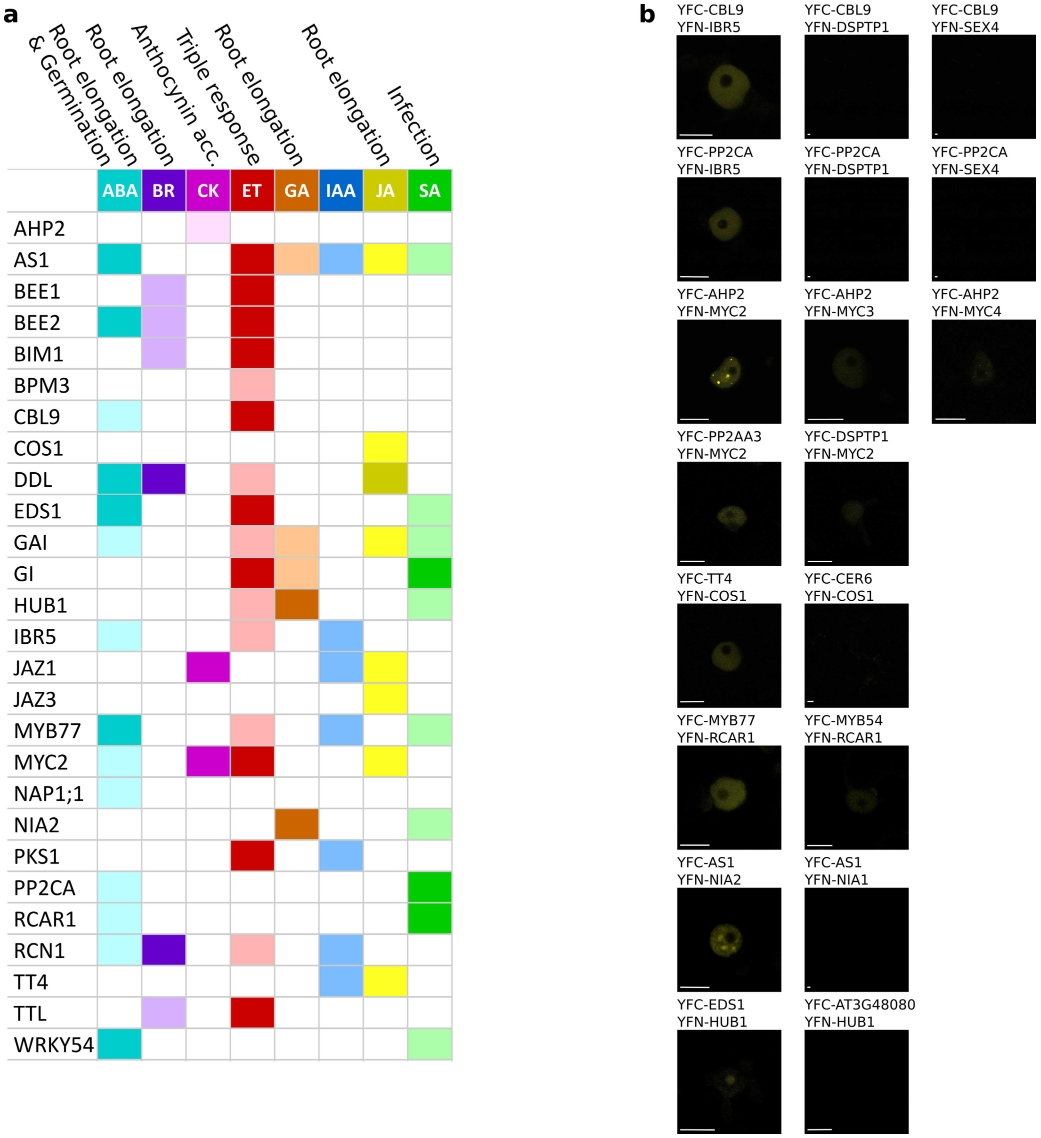
Extended Data Fig. 4 | Hormone-response assays I. a, Rate of ABA-induced germination for Col-0 (wild-type), *cbl9*, *gai*, *myc2*, *ibr5*, *rcar1* and *rcn1* (positive controls), *ahp2*, *as1*, *bee1*, *bee2*, *bim1*, *ddl*, *eds1*, *jaz3*, *myb77* and *wrky54* in the absence (mock) or presence of 0.3 µM ABA. **b**, Root elongation in absence (mock) or presence of 30 µM ABA for the same plant lines as in **a**. **c**, Brassinolide (BL)-mediated inhibition of root length in the absence (mock) or presence of indicated concentrations of BL for Col-0 (wild type), *bee1*, *bee2*, *bim2* (control), *ddl*, *rcn1* and *ttl* (candidate) lines. **d**, Gibberellic acid (GA)-mediated inhibition of root length in the presence of the indicated concentrations of paclobutrazol (Pac) for Col-0 (wild type), *as1*, *gai*, *gi*, *rga*, *rga-28* (control), *hub1*, *jaz3*, *nia2* and *rcn1* (candidate) lines. **b**–**d**, Boxes represent IQRs; bold black lines represent medians; whiskers indicate the highest and lowest data points within 1.5 IQRs; outliers are plotted individually. **a**–**d**, Two-sided *t*-test; **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001. Precise *n* values for each repeat and precise *P* values are provided in Supplementary Table 5.



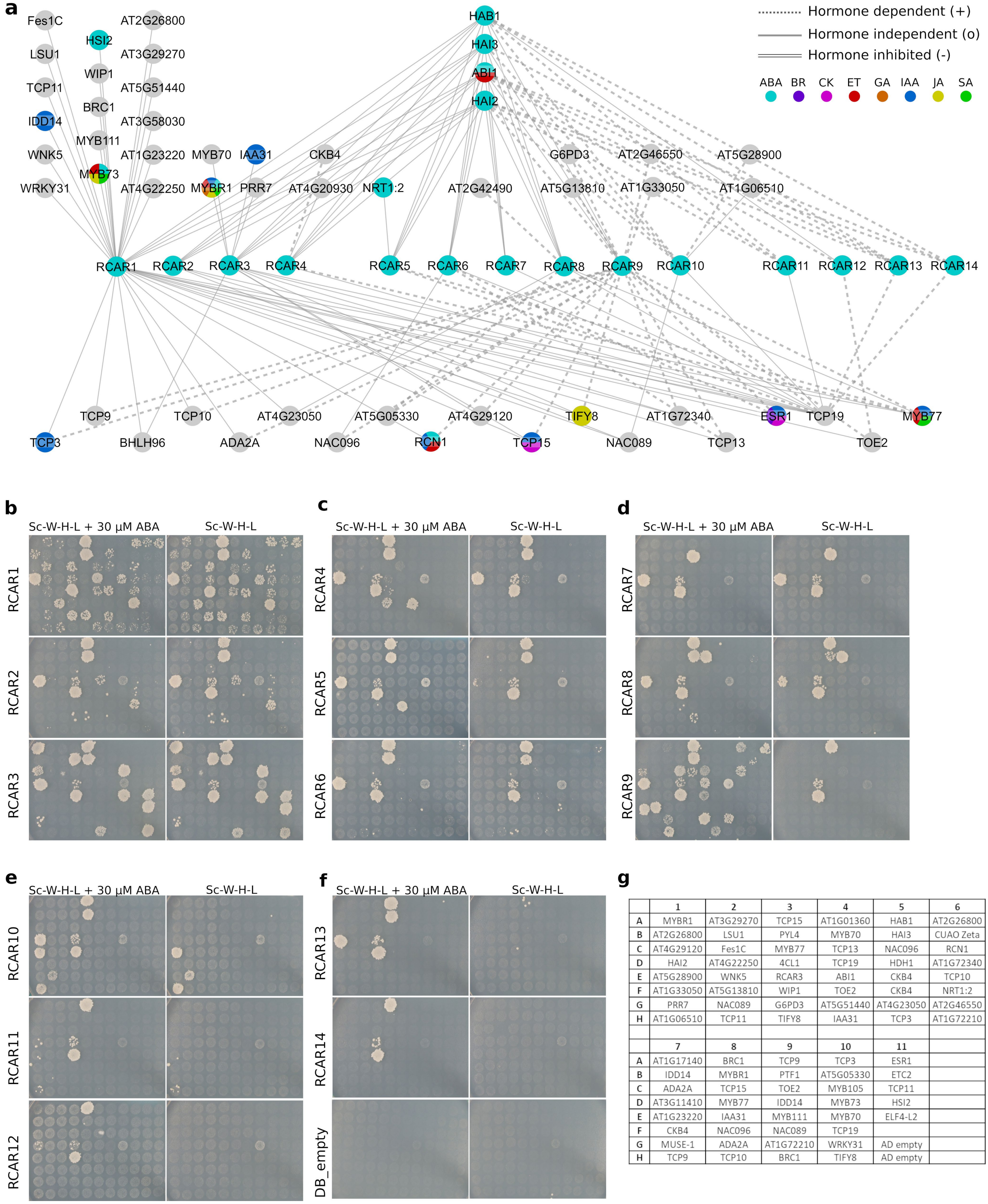
Extended Data Fig. 5 | Hormone-response assays II. a, Salicylic acid (SA)-associated phenotypes, showing *Pst* titres four days after leaves were inoculated with *Pst* by syringe infiltration. In planta *Pst* titres were elevated in mature plants of the indicated genotypes relative to wild-type Col-0 plants. **b**, Jasmonate (JA)-induced root growth in the absence (mock) or presence of 25 µM Me-JA. **c**–**f**, Ethylene-induced triple response in control conditions compared with Col-0 plants. **c**, Apical hook formation in the absence or presence of 10 µM ACC. Shown are representative results underlying the quantitation in **d**. Scale bars, 5 mm. **d**, Proportion of apical loop formation (rather than hook formation) following treatment with 10 µM ACC, for the same lines as in **c**. **e**, Hypocotyl length in the absence or presence of 10 µM ACC for the same lines as in **d**. **f**, Root elongation in the absence or presence of 10 µM ACC for same lines as in **d**. Two-sided *t*-test; *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$. **b**, **e**, **f**, Boxes represent IQRs; black lines represent medians; whiskers indicate highest and lowest data points within 1.5 IQRs; outliers are plotted individually. **a**, **b**, **d**–**f**, Two-sided *t*-test; *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$. Precise *n* values for each repeat and exact *P* values are provided in Supplementary Table 5.



Extended Data Fig. 6 | Ethylene-induced triple response assays (negative controls). The ethylene-mediated triple response in negative control lines is compared with that in Col-0 and *ein3* lines. **a**, Proportion of apical loop formation in response to 10 µM ACC. **b**, Hypocotyl length in the absence or presence of 10 µM ACC. **c**, Root elongation in the absence or presence of 10 µM ACC. Two-sided *t*-test; $^{*}P \le 0.05$, $^{**}P \le 0.01$. **b**, **c**, Boxes represent IQRs; black lines represent medians; whiskers indicate highest and lowest data points within 1.5 IQRs; outliers are plotted individually. Precise *n* values for each repeat and *P* values are in Supplementary Table 5.

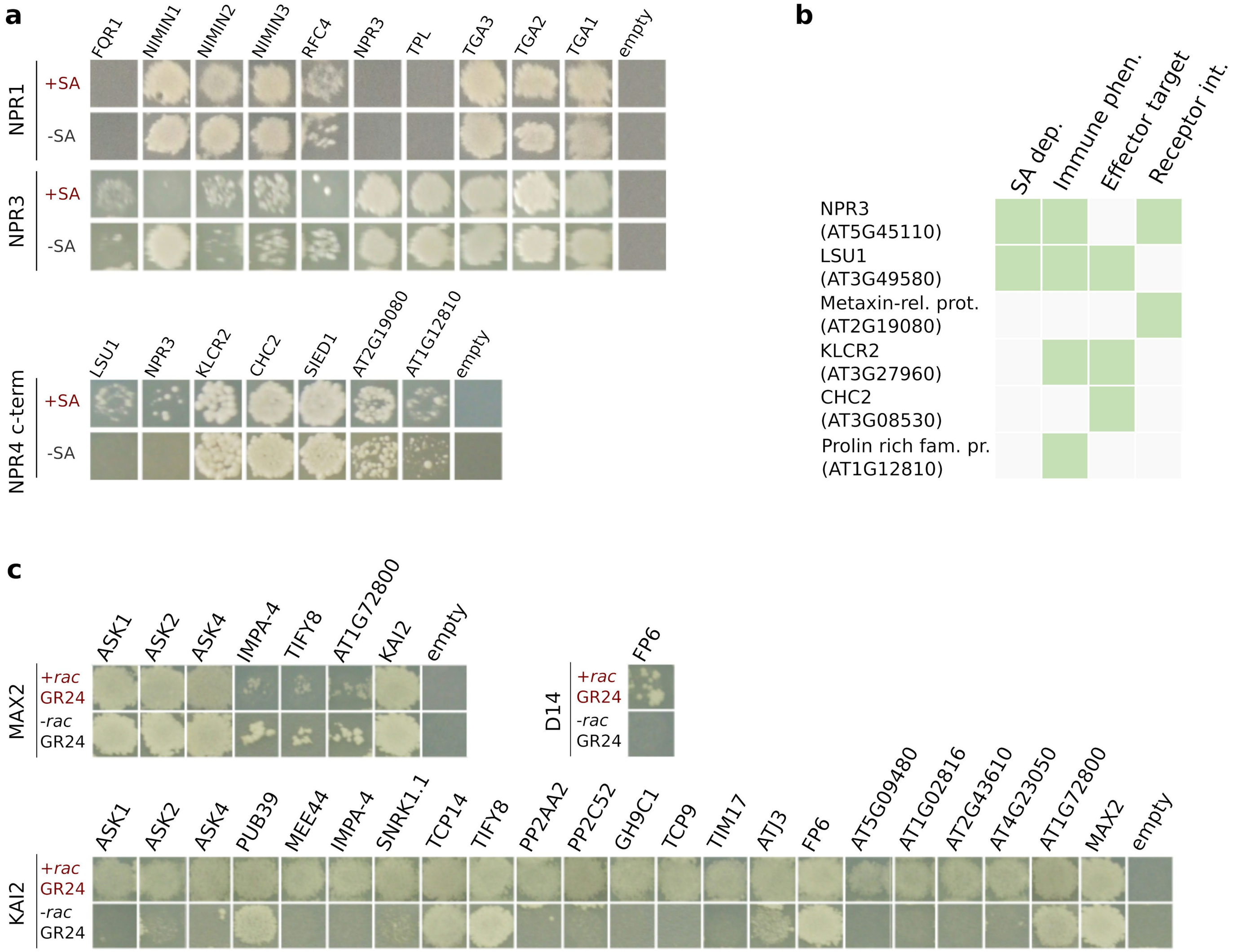


Extended Data Fig. 7 | PCP validation. a, Summary of hormone-assay results for 27 candidate genes. Light colours indicate previously known hormone pathway annotations. Bright colours indicate significant (Fig. 2, Extended Data Figs. 4–7 and Supplementary Table 5) new phenotypes observed in validation assays. **b**, BiFC analysis in *N. benthamiana* of two PCP_I pairs (AHP2–MYC2, MYB77–RCAR1) and five PCP_II pairs (CBL9–IBR5, PP2CA–IBR5, TT4–COS1, AS1–NIA2, EDS1–HUB1). PCP pairs were also tested with one or two negative controls in the BiFC assay. Each construct was tested in duplicate and in two independent assays; one representative result is shown. Scale bars, 10 µm.



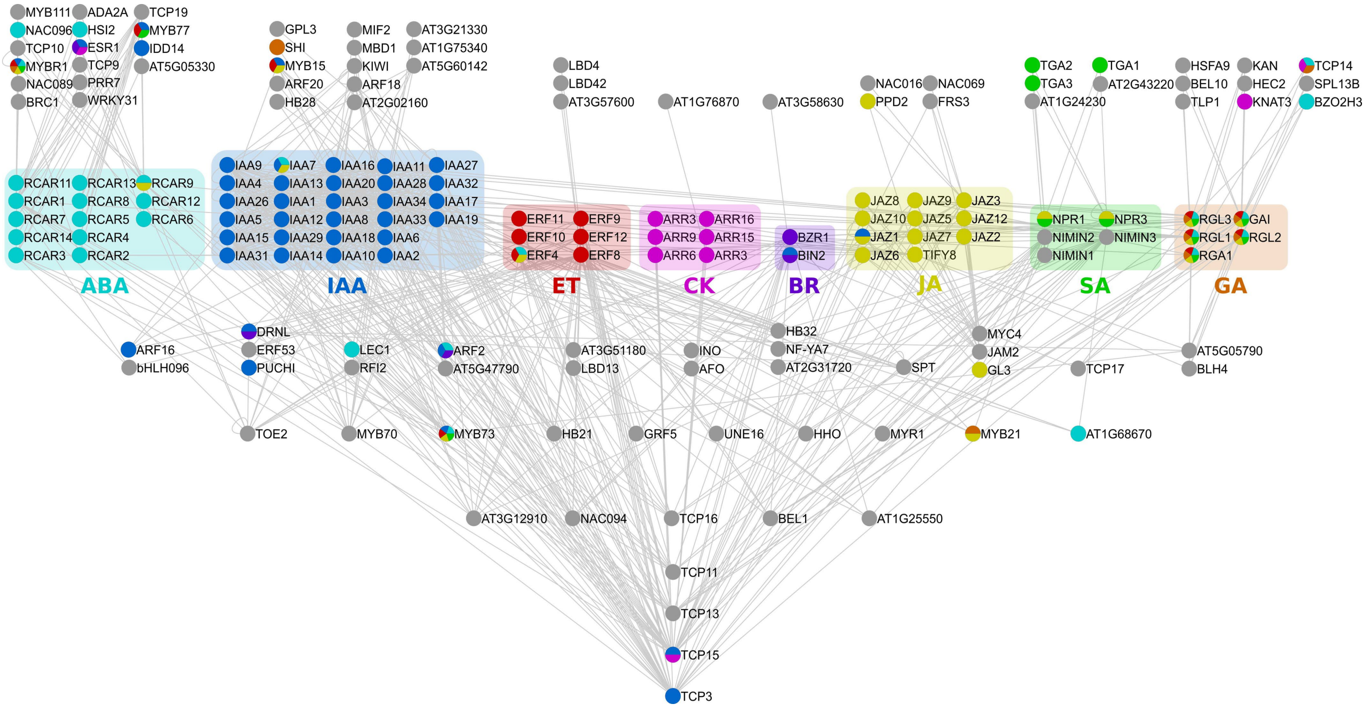
	1	2	3	4	5	6
A	MYBR1	AT3G29270	TCP15	AT1G01360	HAB1	AT2G26800
B	AT2G26800	LSU1	PYL4	MYB70	HAI3	CUAO Zeta
C	AT4G29120	Fes1C	MYB77	TCP13	NAC096	RCN1
D	HAI2	AT4G22250	4CL1	TCP19	HDH1	AT1G72340
E	AT5G28900	WNK5	RCAR3	ABI1	CKB4	TCP10
F	AT1G33050	AT5G13810	WIP1	TOE2	CKB4	NRT1:2
G	PRR7	NAC089	G6PD3	AT5G51440	AT4G23050	AT2G46550
H	AT1G06510	TCP11	TIFY8	IAA31	TCP3	AT1G72210
	7	8	9	10	11	
A	AT1G17140	BRC1	TCP9	TCP3	ESR1	
B	IDD14	MYBR1	PTF1	AT5G05330	ETC2	
C	ADA2A	TCP15	TOE2	MYB105	TCP11	
D	AT3G11410	MYB77	IDD14	MYB73	HSI2	
E	AT1G23220	IAA31	MYB111	MYB70	ELF4-L2	
F	CKB4	NAC096	NAC089	TCP19		
G	MUSE-1	ADA2A	AT1G72210	WRKY31	AD empty	
H	TCP9	TCP10	BRC1	TIFY8	AD empty	

Extended Data Fig. 8 | ABA Y2H interactions. a, ABA-dependent and -independent interactions of the ABA receptors RCAR1–14. All identified interactors were systematically tested against all receptors in the presence or absence of 30 μM ABA. Excluding PP2Cs, single RCAR-specific interactors are displayed above the receptors; interactors common to several RCARs are shown below. Node colours represent hormone annotations. **b**–**f**, One representative set of Y2H results, out of four repeats, showing yeast growth on selective media in the presence or absence of 30 μM ABA. All candidate interactors identified in primary screens were tested systematically against all receptors in the shown representative verification experiments. **g**, Plate layout of candidate interactors tested with the indicated RCARs in **b**–**f**.



	SA dep.	Immune phen.	Effector target	Receptor int.
NPR3 (AT5G45110)	🟩	🟩		🟩
LSU1 (AT3G49580)	🟩	🟩	🟩	
Metaxin-rel. prot. (AT2G19080)				🟩
KLCR2 (AT3G27960)		🟩	🟩	
CHC2 (AT3G08530)			🟩	
Prolin rich fam. pr. (AT1G12810)		🟩		

Extended Data Fig. 9 | Hormone-dependent Y2H interactions. **a**, SA-dependent interactors of NPR1,3,4 in the presence or absence of 100 µM SA. **b**, Evidence for NPR4 interactor functions in plant defence. **c**, MAX2, D14 and KAI2 interactions in the presence or absence of 5 µM rac–GR24. **a**, **c**, One representative set of Y2H results, out of four, showing yeast growth in the presence or absence of hormone. All candidate interactors identified in a primary screen were tested against all receptors in the shown representative verification experiments.



Extended Data Fig. 10 | Pathway convergence on transcription factors. Y2H-derived map showing interactions of repressors and non-DNA-binding transcriptional regulators (boxed and colour-coded for involvement in the relevant main pathway) with *Arabidopsis* transcription factors. Above the repressors/regulators are transcription factors that interact specifically with regulators from one pathway. Lower layers show transcription factors that interact with regulators from several pathways. Node annotations are represented by the indicated colour codes.

natureresearch

Corresponding author(s): Pascal Falter-Braun

Last updated by author(s): 08.4.2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, seeAuthors & Referees and theEditorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data for PRS curation were downloaded from IntAct (August 2014) and BioGRID (version 3.2.115). Data for network comparisons were downloaded from the same databases in June 2016 (IntAct) and Nov 2016 (BioGrid ver 3.4.142). Hormone annotations for all analyses were downloaded from Arabidopsis hormone database version 2.0, and extracted from TAIR10 GO annotations (03/08/2018). Transcriptome data for edge score analysis were taken from the Supplementary material of the referenced paper.
Data analysis	The used public software packages are referred to in the materials and methods section. Custom scripts have been deposited in github as indicated in the code availability statement. Software and packages used for data analysis are Cytoscape 3.7.2, R 3.6.1; Fiji Software Simple Neurite Tracer (v 3.1.3), igraph 1.2.4; drc 3.0-1; topGO 2.36; Kallisto v0.45; FastQC v0.11.7; Trimmomatic v0.36;

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All functional, genetic, and interaction data generated in this study are available as supplementary information. The genes selected for interactome mapping (search space) are presented in Supplementary Information Table 1. All protein-protein interaction data acquired in this study can be found in Supplementary Information Table 2. The data for genetic validation assays can be found in Supplementary Information Table 5. The preliminary edge-scores for all interactions identified in this study are presented in Supplementary Information Table 6. Additionally, all protein interactions from this work have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct5 and assigned the identifier IM-27834.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For all network rewiring and bootstrapping analyses the number of iterations was set to n = 1,000 as this provides a reasonable balance between statstical power and computational effort. For all plant phenotyping experiments we aimed to get a total n of ~100 plants, which provides a good compromise between statistical power and experimental workload and is widely accepted in the community. We started with 40-50 plants for each of three repeats which were germinated in standard conditions. Subsequently, plants at a similar developmental stage were selected for transfer to the hormone or mock-control plates, respectively. The selection of similar developmtal stages is important to ensure comparability of results and exclude developmental stage as a confounding factor. Due to stochastic biological variabilty this selection led to moderately variable number of plants in each experiment, i.e. a variable n in each assay and repeat. For the germination assay, we also had the limitation that the seeds needed to fit onto a smaller plate that fit under the microscope, which is why the numbers are lower. For all plant assays the n of each repeat and each line is provided in the extended data tables.
Data exclusions	No data were excluded. During the phenotyping experiments, individual seedlings at similar developmental stage were selected to exlude developmental stage as a confounding variable. This was prior to observation of the outcome and even prior to start of treatment.
Replication	As part of the interaction mapping pipeline all Y2H interactions are verified at least three times. All functional genetic assays have at least three technical and three biological replicates. Cell biology data were repeated at least once (two repeats). For phenotypic assays we aimed for three, or two biological repeats.
Randomization	Network statistics are always checked against randomized networks to exclude pure effects due to network topology that may be independent of the measured biology.
Blinding	All in planta hormone validation assays, the yeast benchmarking using PRS/RRS and the validation assays were done and scored in a randomized design in which the experimenters were unaware of the identity of the samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Only commercial anti-tag antibodies were used. Specifically M2 anti-Flag (Sigma), anti-GST (Amersham Biosciences) and anti-MBP (New England Biolabs)
Validation	We mostly relied on the quality control and many years of experience of the scientific community with these standard regeants. We then only verified that the antibodies recognized the tags they were raised against. For the anti-Flag pull-down experiments negative controls were included to ensure specificity of the purification.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	none, only plants from stock centers - numbers are provided in Supplementary Table 5
Wild animals	none, only plants from stock centers - numbers are provided in Supplementary Table 5
Field-collected samples	none, only plants from stock centers - numbers are provided in Supplementary Table 5
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Transcriptional regulation of strigolactone signalling in *Arabidopsis*

https://doi.org/10.1038/s41586-020-2382-x

Received: 3 February 2019

Accepted: 24 March 2020

Lei Wang[1,2,6], Bing Wang[1,6]✉, Hong Yu[1], Hongyan Guo[1], Tao Lin[1], Liquan Kou[1], Anqi Wang[1,3], Ning Shao[1], Haiyan Ma[1], Guosheng Xiong[2,4], Xiaoqiang Li[5], Jun Yang[5], Jinfang Chu[1,3] & Jiayang Li[1,3]✉

Plant hormones known as strigolactones control plant development and interactions between host plants and symbiotic fungi or parasitic weeds[1–4]. In *Arabidopsis thaliana* and rice, the proteins DWARF14 (D14), MORE AXILLARY GROWTH 2 (MAX2), SUPPRESSOR OF MAX2-LIKE 6, 7 and 8 (SMXL6, SMXL7 and SMXL8) and their orthologues form a complex upon strigolactone perception and play a central part in strigolactone signalling[5–10]. However, whether and how strigolactones activate downstream transcription remains largely unknown. Here we use a synthetic strigolactone to identify 401 strigolactone-responsive genes in *Arabidopsis*, and show that these plant hormones regulate shoot branching, leaf shape and anthocyanin accumulation mainly through transcriptional activation of the *BRANCHED 1*, *TCP DOMAIN PROTEIN 1* and *PRODUCTION OF ANTHOCYANIN PIGMENT 1* genes. We find that SMXL6 targets 729 genes in the *Arabidopsis* genome and represses the transcription of *SMXL6*, *SMXL7* and *SMXL8* by binding directly to their promoters, showing that SMXL6 serves as an autoregulated transcription factor to maintain the homeostasis of strigolactone signalling. These findings reveal an unanticipated mechanism through which a transcriptional repressor of hormone signalling can directly recognize DNA and regulate transcription in higher plants.

Strigolactones are a class of carotenoid-derived plant hormones[1,2] that have fundamental effects on shoot branching, leaf development, plant height, anthocyanin accumulation, root architecture, and adaptation to drought and phosphate starvation[3,4]. The D14 protein in *Arabidopsis*[5], rice[6] or pea[11] and orthologues of the KARRIKIN INSENSITIVE2 (KAI2) protein in *Striga hermonthica*[12–14] have been identified as strigolactone receptors, which hydrolyse strigolactones, bind covalently to intermediate molecules, and form a complex with the F-box protein D3 and the transcriptional repressor D53 (refs. [5–8]). D3 can adopt two conformations to regulate the hydrolysis activity of D14 and to mediate the ubiquitination and degradation of D53 (ref. [6]). D53 and its orthologues SMXL6, SMXL7 and SMXL8 (SMXL6, 7, 8 hereafter) in *Arabidopsis* have been proposed to be transcriptional repressors that function by interacting with transcription factors and with the transcriptional corepressor proteins TOPLESS (TPL) and TPL RELATED (TPR)[4,7–10,15].

Enormous efforts have been made to characterize strigolactone-responsive genes, but only a few have been identified so far, including *BRANCHED 1* (*BRC1*) in *Arabidopsis* and pea[16,17], *MAX3*, *MAX4*, *SMXL2*, *6*, *7*, *8*, *SALT TOLERANCE HOMOLOG 7*/*BZR1-1D SUPPRESSOR 1* (*STH7*/*BZS1*), *DWARF14-LIKE2* (*DLK2*), *KAR-UP F-BOX1* (*KUF1*) and several auxin-responsive genes in *Arabidopsis*[10,16,18–20], and *D53* and *CYTOKININ OXIDASE/DEHYDROGENASE9* (*OsCKX9*) in rice[7,8,21]. These modest transcriptional changes cannot explain how strigolactones regulate diverse aspects of plant development and responses to various environmental signals. Furthermore, the synthetic strigolactone analogue *rac*-GR24, widely used experimentally, comprises a pair of enantiomers, $\mathrm{GR24}^{5\mathrm{DS}}$ and $\mathrm{GR24}^{ent\text{-}5\mathrm{DS}}$ (Extended Data Fig. 1a), and stimulates both the strigolactone and the karrikin signalling pathways[4,18]. It is therefore crucial to explore strigolactone-responsive genes by using chemicals that specifically stimulate strigolactone signalling with high efficiency.

Here we report the identification of key strigolactone-responsive genes that specifically regulate various developmental processes such as shoot branching, leaf development, anthocyanin accumulation, and drought adaption. More importantly, we further show that SMXL6 can directly bind to the promoters of *SMXL6*, *7*, *8* and regulate their transcription, thus functioning as a transcription factor in strigolactone signalling.

Identifying strigolactone-responsive genes

To identify genes that respond specifically to strigolactones in *Arabidopsis*, we compared the effects of *rac*-GR24 with those of synthetic $\mathrm{GR24}^{5\mathrm{DS}}$, $\mathrm{GR24}^{4\mathrm{DO}}$ and $\mathrm{GR24}^{ent\text{-}4\mathrm{DO}}$ (Extended Data Fig. 1a). We found that $\mathrm{GR24}^{4\mathrm{DO}}$ could trigger the degradation of SMXL6 labelled with green fluorescent protein (GFP), but that the enantiomer $\mathrm{GR24}^{ent\text{-}4\mathrm{DO}}$ had few effects (Extended Data Fig. 1b). Although 5DS stereoisomers are

---

[1]State Key Laboratory of Plant Genomics, and National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing, China. [2]Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [3]University of Chinese Academy of Sciences, Beijing, China. [4]Plant Phenomics Research Center, Nanjing Agricultural University, Nanjing, China. [5]CAS Key Laboratory of Energy Regulation, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China. [6]These authors contributed equally: Lei Wang, Bing Wang. ✉e-mail: bingwang@genetics.ac.cn; jyli@genetics.ac.cn

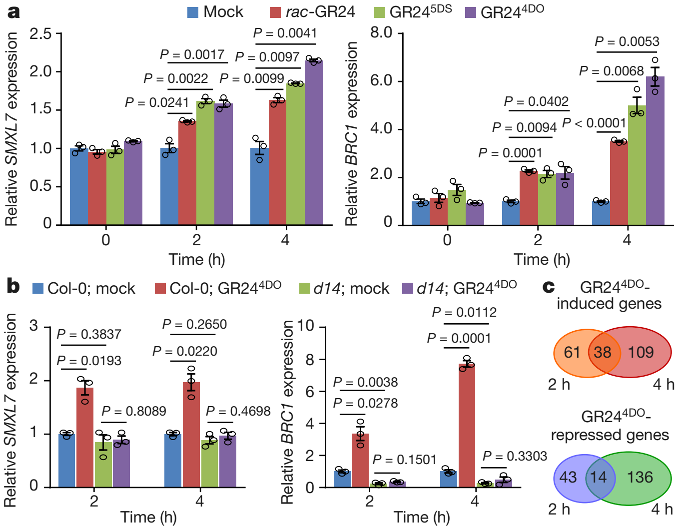


Fig. 1 | GR24^{4DO} efficiently stimulates strigolactone signalling. a, Expression of *SMXL7* and *BRC1* in 10-day-old wild-type seedlings treated with 5 µM *rac*-GR24, GR24^{5DS} or GR24^{4DO}. Data are means ± s.e.m.; $n = 3$ biologically independent samples. **b**, Induction of *SMXL7* and *BRC1* upon GR24^{4DO} treatment was blocked by *D14* mutation (*d14* plants). Data are means ± s.e.m.; $n = 3$ biologically independent samples. Col-0 plants are wild type. **c**, Overview of genes upregulated or downregulated upon treatment with 5 µM GR24^{4DO} for 2 h or 4 h in three RNA-seq replicates. **a**, **b**, *P* values are shown; two-sided Student's *t*-test.

more effective than 4DO stereoisomers in triggering the strigolactone biosensor StrigoQuant[22], GR24^{5DS} can activate both strigolactone and karrikin signalling through D14 and KAI2 (ref. [23]). We found that GR24^{4DO} promoted the expression of *SMXL6*, *7*, *8* and *BRC1* with higher efficiency than GR24^{5DS} in a D14-dependent manner (Fig. 1a, b and Extended Data Fig. 1c, d), indicating that GR24^{4DO} specifically induces transcriptional responses[18] through D14 and is an ideal chemical for experimental stimulation of strigolactone signalling.

We further conducted RNA-sequencing (RNA-seq) analyses upon GR24^{4DO} treatment, and identified 99 upregulated and 57 downregulated genes after 2 hours, as well as 147 upregulated and 150 downregulated genes after 4 hours (Fig. 1c and Supplementary Table 1). These differentially expressed genes (DEGs) cover nearly a quarter of the genes that have previously been reported to be responsive to GR24^{5DS} and *rac*-GR24 (Supplementary Table 1). More importantly, roughly 90% of DEGs were newly found (Supplementary Table 1) and could be verified by reverse transcription with quantitative polymerase chain reaction (RT–qPCR; Extended Data Fig. 2), indicating a high efficiency of GR24^{4DO} in stimulating strigolactone signalling in *Arabidopsis*.

The most highly enriched gene ontology (GO) terms in the DEGs upregulated at 2 hours were involved in microtubule function (Supplementary Table 2), suggesting a potential regulation of the cytoskeleton by strigolactone signalling. Early auxin-inducible genes were enriched in the 2-hour downregulated DEGs (Supplementary Table 2 and Extended Data Fig. 2a), consistent with the previous finding that most of the *rac*-GR24 repressed genes are induced by auxins[16]. Upon 4 hours of treatment, expression of the auxin-biosynthesis genes *YUCCA 3* (*YUC3*) and *YUC5* was downregulated (Supplementary Table 1 and Extended Data Fig. 2b). Strigolactones have previously been found to induce clathrin-mediated endocytosis of PIN-FORMED 1 (PIN1) to deplete membrane-localized PIN1 and repress polar auxin transport in the stem[24], while auxin promotes expression of the strigolactone-biosynthesis genes *MAX3* and *MAX4* (ref. [25]). These results indicate complex crosstalk between strigolactones and auxin, both of which are important in plant architecture.

We found that GR24^{4DO} induces the expression of *At14a-LIKE 1* (*AFL1*; Extended Data Figs. 2c, 3a and Supplementary Table 1), which is required for drought tolerance in *Arabidposis*[26]. More importantly, the *smxl6*, *smxl7*, *smxl8* triple mutant (*s678* hereafter) displayed strong drought tolerance and elevated *AFL1* expression (Extended Data Fig. 3b, c). Several karrikin-inducible genes[19,27] were also promoted by GR24^{4DO} (Extended Data Fig. 2d and Supplementary Tables 1, 2), suggesting crosstalk between the strigolactone and karrikin signalling pathways[18]. In addition, analysis of the Kyoto Encyclopedia of Genes and Genomes (KEGG) showed that carotenoid and flavonoid biosynthesis pathways were enriched in 2-hour and 4-hour upregulated DEGs, respectively (Supplementary Table 2 and Extended Data Fig. 2e).

To investigate how strigolactones control plant development through transcriptional regulation of responsive genes, we systematically analysed DEGs that encode transcription factors, finding 24 upregulated and 14 downregulated genes (Extended Data Fig. 2f, g and Supplementary Table 3). We then focused on *BRC1*, *TCP1* and *PRODUCTION OF ANTHOCYANIN PIGMENT 1* (*PAP1*) to study their roles in strigolactone-regulated plant development.

Strigolactones, *BRC1* and shoot branching

SMXL6 represses transcription by interacting with TPL/TPR proteins in a manner that depends on its ethylene-responsive-element-binding-factor-associated amphiphilic repression (EAR) motif[7–10]. We found that wild-type *SMXL6* could rescue the shoot-branching phenotype of the *s678* triple mutant, whereas the EAR-deleted *SMXL6* (*SMXL6*$^{\Delta EAR}$) could not (Extended Data Fig. 4a, b), indicating that transcriptional regulation through the EAR motif is required for strigolactone-mediated shoot branching. BRC1 is a key regulator that represses bud outgrowth and functions as a signal integrator in numerous pathways[25]. In unelongated axillary buds, *BRC1* expression is dramatically repressed by SMXL6 (refs. [9,10]) in an EAR-dependent manner (Extended Data Fig. 5a). The high-branching mutant *brc1-6*, a null allele of *BRC1*, completely suppressed the shoot-branching phenotypes of *s678* (Extended Data Fig. 5b–f), indicating an essential role of BRC1 in strigolactone-inhibited shoot branching[25,28]. Furthermore, GR24^{4DO} induced the expression of *HB40*—a BRC1-target gene that regulates the biosynthesis of abscisic acid (ABA)[29]—in a manner dependent on *D14* and *BRC1* (Fig. 2a and Extended Data Fig. 5g). Notably, endogenous ABA levels in unelongated axillary buds were reduced in the strigolactone-biosynthesis-deficient mutant *max3-9* but elevated in *s678* plants, and this elevation could be repressed by *brc1-6* (Fig. 2b, Extended Data Fig. 5h), indicating that strigolactones elevate ABA levels in buds through transcriptional induction of *BRC1*. These results indicate that the EAR motif of SMXL6 is essential for strigolactone-activated *BRC1* expression in axillary buds, which promotes ABA accumulation and represses bud outgrowth. The shoot auxin-transport system is also important in strigolactone-regulated shoot branching, probably in an EAR-independent manner[28,30,31]. Thus, SMXL6, 7 and 8 promote shoot branching through transcriptional repression of *BRC1* and nontranscriptional regulation of auxin transport.

Strigolactones, *TCP1* and leaf shape

Strigolactones promote leaf elongation through SMXL6 in an EAR-dependent manner (Extended Data Fig. 4c, d). However, the leaf shape of the *s678 brc1-6* quadruple mutant was similar to that of the *s678* mutants (Extended Data Fig. 6a), raising the possibility that strigolactones also regulate leaf development through other downstream genes. We found that the expression of *TCP1*, which regulates leaf development[32], was induced upon GR24^{4DO} treatment in a D14-dependent manner (Fig. 2c). The *TCP1* expression level was greatly repressed in *max3-9* but dramatically increased in *s678* and *max3-9 s678* mutants (Fig. 2d), and the transcriptional regulation of *TCP1* depended on the EAR motif of SMXL6 (Extended Data Fig. 6b). Furthermore, overexpression of TCP1–SRDX—a dominant-negative form of TCP1 with an added 12-amino-acid repressor sequence[32]—resulted in rounder leaves in

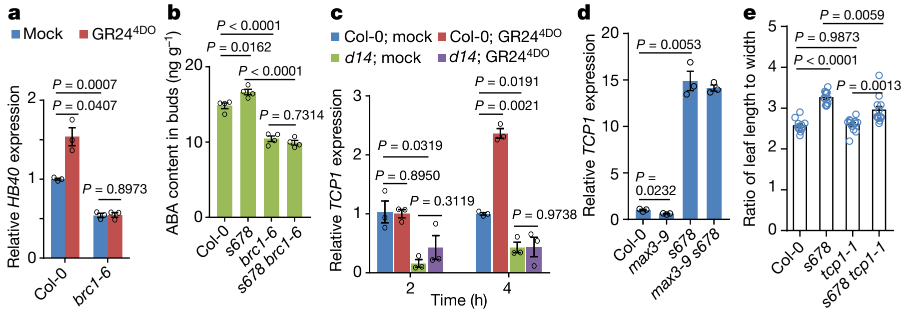


Fig. 2 | *BRC1* and *TCP1* regulate shoot branching and leaf shape, respectively, in strigolactone signalling. a, Expression of *HB40* after treatment with GR24^{4DO} for 4 h in wild-type (Col-0) and *brc1-6* plants. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **b**, ABA content in unelongated axillary buds of wild-type and the indicated mutant plants. Data are means ± s.e.m.; *n* = 4 biologically independent samples. **c**, Expression of *TCP1* in wild-type and *d14-1* seedlings treated with GR24^{4DO}. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **d**, Expression of *TCP1* in wild-type and mutant seedlings. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **e**, Morphology of the fifth leaves of 3-week-old wild-type and indicated mutant plants. Data are means ± s.e.m.; *n* = 12 leaves. **a**–**e**, *P* values are shown; Tukey's honest significant difference (HSD) test (**b**, **e**) or two-sided Student's *t*-test (**a**, **c**, **d**).

both wild-type and *s678* backgrounds (Extended Data Fig. 6c). Notably, although the *tcp1-1* null mutant formed leaves similarly to the wild type, it greatly reduced the ratio of leaf length to width in the *s678* background (Fig. 2e and Extended Data Fig. 6d, e), indicating that *SMXL6*, *7*, *8* suppress leaf elongation partially by repressing *TCP1* expression. Therefore, *TCP1* makes an important contribution to leaf-shape regulation in strigolactone signalling.

Strigolactones, PAPs and anthocyanin synthesis

Transcripts of *PAP1*, *PAP2*, *MYB113* and *MYB114*—which encode R2R3-MYB transcription factors and activate anthocyanin biosynthesis—were greatly induced after GR24^{4DO} treatment for 2 hours and 4 hours (Fig. 3a and Extended Data Fig. 7a). Furthermore, after GR24^{4DO} treatment for 4 hours, expression of the anthocyanin-biosynthesis genes *DIHYDROFLAVONOL 4-REDUCTASE* (*DFR*), *TRANSPARENT TESTA 7* (*TT7*) and *ANTHOCYANIDIN SYNTHASE* (*ANS*), which are activated by PAP1, PAP2, MYB113 and MYB114 (ref. [33]), was induced (Fig. 3a and Extended Data Fig. 2e). Endogenous anthocyanin levels were moderately decreased in *max3-9* but greatly increased in *s678* mutants (Extended Data Fig. 7b, c). Accordingly, expression levels of *PAP1*, *PAP2*, *MYB113*, *MYB114* and *DFR* were reduced in *max3-9* but dramatically increased in *s678* and *max3-9 s678* mutants (Extended Data Fig. 7d). Overexpression of *SMXL6* repressed the elevated expression of *PAP1*, *PAP2*, *MYB113*, *MYB114* and *DFR* and anthocyanin accumulation in *s678* in an EAR-dependent manner (Extended Data Fig. 7c, e). Furthermore, the impaired anthocyanin accumulation seen in *max3-9* mutants was completely suppressed by the *pap1-D* mutation (Fig. 3b, c and Extended Data Fig. 7f), which causes constitutive overexpression of *PAP1* and overaccumulation of anthocyanin[33]. More importantly, anthocyanin accumulation in *s678* mutants was partially rescued by the *pap2-1* null mutation, and could be completely rescued by the *pap1-2* null mutation or by *pap1-2 pap2-1* double mutations (Fig. 3d and Extended Data Fig. 7g), showing that PAP1 and PAP2 work downstream of SMXL6, 7, 8 in strigolacton-promoted anthocyanin biosynthesis. The induction of *DFR* expression upon GR24^{4DO} treatment was disrupted in *pap1-D* mutants (Fig. 3e), indicating an important role of *PAP1* in strigolactone-induced *DFR* expression. Together, these results suggest a gene-regulation cascade in which strigolactones induce the expression of *PAP1*, *PAP2*, *MYB113* and *MYB114* and consequently activate the transcription of *DFR*, *ANS* and *TT7* to elevate anthocyanin abundance.

SMXL6, 7, 8 function as transcription factors

To understand how SMXL6 regulates the expression of target genes, we investigated its global binding profiles through chromatin immunoprecipitation sequencing (ChIP–seq) assays in the *s678* mutant and in a transgenic line that overexpresses haemagglutinin-tagged SMXL6 (*pSMXL6:SMXL6-HA*), which rescues the phenotypes and gene-expression profiles of *s678* (Extended Data Figs. 4, 5a, 6b and 7c, e). The two biological replicates shared 1,079 peaks, which were highly enriched in 3-kilobase promoters (59.3%) compared with intergenic (24.3%) and intragenic (16.4%) regions (Fig. 4a, b and Supplementary Tables 4, 5). Among the 729 genes targeted by SMXL6–HA in two replicates, 28 are GR24^{4DO}-responsive genes, leading to roughly 3.22-fold enrichment compared with nonspecific binding (Fisher's exact test, $P = 1.93 \times 10^{-7}$; Supplementary Tables 6, 7). SMXL6–HA could bind the

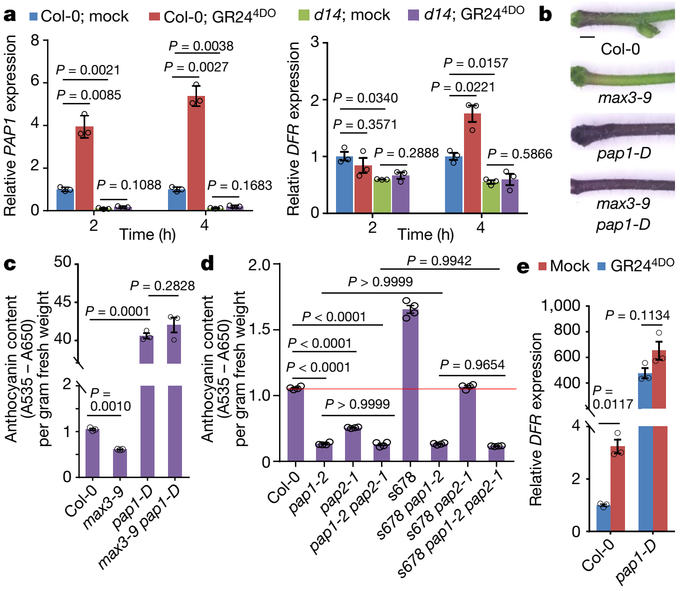


Fig. 3 | Strigolactones promote anthocyanin accumulation by inducing transcription of *PAP1* and anthocyanin-biosynthesis genes. a, Expression of *PAP1* and *DFR* in wild-type (Col-0) and *d14-1* seedlings after GR24^{4DO} treatment. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **b**, Anthocyanin accumulation in the stem base of 3-week-old wild-type and indicated mutant plants. Scale bar, 1 mm. **c**, **d**, Anthocyanin content (A535 and A650 anthocyanins) in 3-week-old wild-type and indicated mutant plants. Data are means ± s.e.m.; *n* = 3 (**c**) or 4 (**d**) pools (6 seedlings per pool). **e**, Expression of *DFR* in wild-type and *pap1-D* seedlings after GR24^{4DO} treatment for 4 h. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **a**, **c**–**e**, *P* values are shown; two-sided Student's *t*-test (**a**, **c**, **e**) or Tukey's HSD test (**d**).

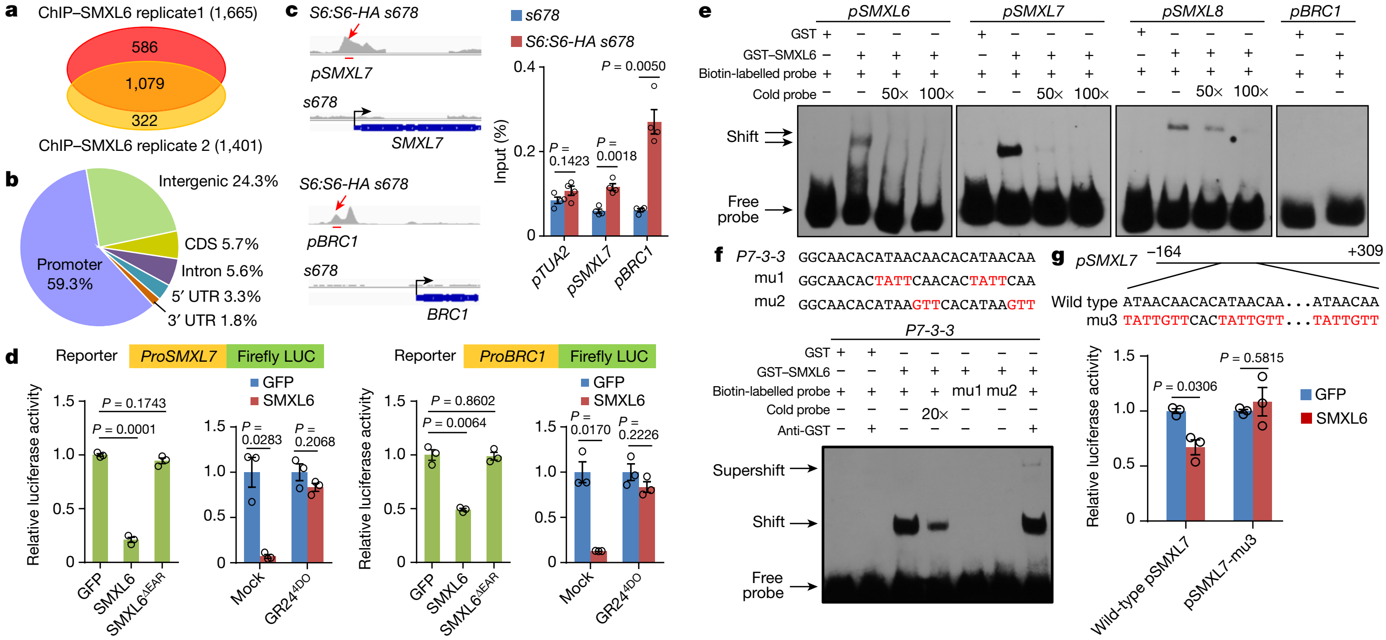


Fig. 4 | SMXL6 binds directly to the *SMXL7* promoter. a, SMXL6-binding sites in two ChIP–seq replicates. **b**, Distribution of overlapping peaks bound by SMXL6 in the *Arabidopsis* genome. CDS, coding sequence. **c**, SMXL6 can associate with the *SMXL7* and *BRC1* promoters in ChIP–seq (left) and ChIP–qPCR assays (right). Chromatin from seedlings of *S6:S6-HA s678* and *s678* plants was immunoprecipitated with anti-haemagglutinin (HA) polyclonal antibodies. Red lines below the peak regions represent probes used in ChIP–qPCR assays (**c**) and EMSAs (**e**). In ChIP–qPCR assays, the enrichment of target gene promoters is displayed as the percentage of input DNA. Data are means ± s.e.m.; *n* = 4 biologically independent samples. The promoter of *TUBULIN ALPHA 2* (*TUA2*) was used as a nonspecific target. **d**, Transcriptional activities of SMXL6 and SMXL6$^{\Delta EAR}$ on the *SMXL7* (left) and *BRC1* (right) promoters in *Arabidopsis* protoplasts. Data are normalized to samples expressing GFP and are means ± s.e.m.; *n* = 3 biologically independent samples. LUC, luciferase. **e**, SMXL6 binds directly to the *SMXL6*, *7* and *8* promoters but not the *BRC1* promoter in EMSAs. For competition, 50- and 100-fold excess unlabelled probes (cold probes) were mixed with biotin-labelled probes. GST, glutathione-*S*-transferase. Data represent three independent experiments. **f**, Fine mapping of the *SMXL7* promoter regions bound by SMXL6. We used 20-fold excess unlabelled probes for competition. The 'supershift' band was detected in the presence of an anti-GST monoclonal antibody. Data represent four independent experiments. **g**, Repression by SMXL6 on the *SMXL7* promoter depends on the ATAACAA motif. Data are normalized to samples expressing GFP. Data are means ± s.e.m.; *n* = 3 biologically independent samples. **c**, **d**, **g**, *P* values are shown; two-sided Student's *t*-test.

promoter regions of *SMXL6*, *7*, *8* and *BRC1*, but no SMXL6-binding signals were found in *TCP1*, *PAP1*, *PAP2*, *MYB113* or *MYB114* (Fig. 4c, Extended Data Fig. 8a and Supplementary Tables 5, 6). Furthermore, transcription of luciferase reporter genes driven by the promoters of *SMXL6*, *7*, *8* and *BRC1* was repressed by SMXL6, and deletion of the EAR motif or GR24^{4DO} treatment substantially released such transcriptional repression (Fig. 4d and Extended Data Fig. 8b, c), indicating that transcriptional regulation of SMXL6 is controlled by strigolactones.

Expression of *SMXL6*, *7*, *8* was dramatically decreased in nonelongated buds of strigolactone-biosynthetic and -signalling mutants, indicating that *SMXL6*, *7*, *8* undergo negative feedback regulation (Extended Data Fig. 8d, e). The finding that a SMXL homologue, HEAT SHOCK PROTEIN 101 (HSP101), can bind to specific messenger RNAs[34] prompted us to examine whether SMXLs have potential DNA-binding activities. Surprisingly, we found that SMXL6 and SMXL7 could bind directly to the promoters of *SMXL6*, *7*, *8*, and that SMXL8 bound directly to the *SMXL7* promoter in vitro, with the shifted band showing a much stronger signal for the *SMXL7* promoter in electrophoretic mobility shift assays (EMSAs; Fig. 4e and Extended Data Fig. 9a, b). SMXL6 could directly bind to the *P7-3* and *P7-3-3* fragments of the *SMXL7* promoter, with either the ATAA to TATT or the CAA to GTT mutation in these fragments completely disrupting the interaction (Fig. 4f and Extended Data Fig. 9c, d). Notably, the binding of SMXL6 to the ATAACAA motif of the *SMXL7* promoter had functional consequences: mutations in this signature greatly disturbed the transcriptional repression activity of SMXL6 in vivo (Fig. 4g), indicating that the ATAACAA motif is essential for SMXL6 to directly bind and inhibit *SMXL7*. Consistent with this result, ATAACAA is 1.66-fold and 3.31-fold enriched in, respectively, 1-kb and 100-bp flanking sequences around the ChIP–seq peak summits (*P* value less than 1×10^{-5} by χ^2 test) of GR24^{4DO}-responsive genes (Supplementary Table 7) compared with the 1-kb promoter of all genes in the *Arabidopsis* genome. In addition, SMXL6 could not directly bind to the *BRC1* promoter in EMSAs (Fig. 4e), suggesting that SMXL6 may work together with unknown transcription factors to repress *BRC1* expression. Together, these results show that the SMXL6, 7, 8 proteins can bind DNA directly and negatively regulate their own transcription, functioning as autoregulated transcription factors to maintain the homeostasis of SMXL6, 7, 8 and to downregulate strigolactone signalling. Our findings reveal a previously unknown mechanism for transcriptional regulation in plant hormone signalling pathways.

Discussion

Here we have demonstrated an autoregulation model for strigolactone signalling (Fig. 5). SMXL6 binds directly to the promoters of *SMXL6*, *7*, *8* and negatively regulates their transcription; strigolactones trigger the formation of the SMXL6–D14–MAX2 complex and degradation of SMXL6 through the ubiquitination proteolysis system, thus releasing the transcriptional inhibition of *SMXL6*, *7*, *8* and forming a negative feedback loop that is required for the strict modulation of SMXL6, 7, 8 abundance. Meanwhile, SMXL6 can also, together with the TPL/TPR proteins, function as a transcriptional repressor to regulate shoot branching, leaf elongation and anthocyanin biosynthesis, mainly by repressing the transcription of *BRC1*, *TCP1* and *PAP1*, respectively (Fig. 5). These findings show a panorama of transcriptional regulation in strigolactone signalling. Further identifying and investigating the

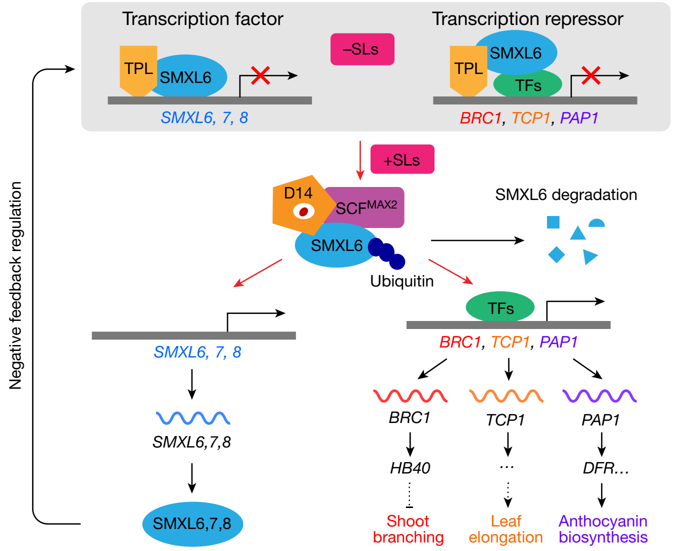


Fig. 5 | Proposed model of transcriptional regulation by SMXL6. In the absence of strigolactones (SLs; top), SMXL6 plus TPL bind directly to the promoters of *SMXL6*, *7*, *8* and repress their expression, functioning as a repressive transcription factor (TF). Meanwhile, SMXL6 can also form a complex with unknown transcription factors that are expected to recognize and bind to the promoters of *BRC1*, *TCP1* or *PAP1*, repressing their transcription as well. In the presence of SLs, D14 binds SLs (white and red circle within D14) and promotes the formation of the D14–$\mathrm{SCF}^{\mathrm{MAX2}}$–SMXL6 complex, triggering the ubiquitin-mediated degradation of SMXL6. This relieves the transcriptional repression of *SMXL6*, *7*, *8*. Newly synthesized SMXL6 proteins in turn repress transcription, forming a negative feedback loop. The degradation of SMXL6 also releases its transcriptional repression of *BRC1*, *TCP1* and *PAP1*, thus activating signalling cascades that repress shoot branching, promote leaf elongation and enhance anthocyanin biosynthesis, respectively.

transcription factors associated with SMXL6 and their modulation of early-responsive genes will uncover how strigolactones regulate diverse aspects of plant development and symbiotic relationships with arbuscular mycorrhiza fungi.

In rice, IDEAL PLANT ARCHITECTURE 1 (IPA1)—a key transcription factor that controls plant architecture—can interact physically with D53 and function in the feedback regulation of strigolactone-induced *D53* expression[15]. The feedback regulation involving SMXL6 is different, however. Furthermore, *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE9* (*SPL9*) and *SPL15*, the orthologues of *IPA1* in *Arabidopsis*, function in parallel pathways with *SMXL6*, *7*, *8* in the regulation of shoot branching (Extended Data Fig. 10), indicating diverse signalling mechanisms in monocotyledonous and dicotyledonous plants[15,35]. Our finding that SMXL6 can bind DNA and functions as a transcription factor distinguishes strigolactone signalling from other plant hormone signalling pathways, including the auxin, jasmonate and gibberellin pathways, that feature degradation mediated by the SCF ubiquitin ligase, thus revealing noncanonical mechanisms across phytohormone signalling. Re-evaluation of the direct targets of essential transcriptional repressors in hormone signalling promises to reveal hitherto hidden mechanisms underlying plant development, evolution and environmental adaptation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2382-x.

1. Gomez-Roldan, V. et al. Strigolactone inhibition of shoot branching. *Nature* **455**, 189–194 (2008).
2. Umehara, M. et al. Inhibition of shoot branching by new terpenoid plant hormones. *Nature* **455**, 195–200 (2008).
3. Al-Babili, S. & Bouwmeester, H. J. Strigolactones, a novel carotenoid-derived plant hormone. *Annu. Rev. Plant Biol*. **66**, 161–186 (2015).
4. Waters, M. T., Gutjahr, C., Bennett, T. & Nelson, D. C. Strigolactone signaling and evolution. *Annu. Rev. Plant Biol*. **68**, 291–322 (2017).
5. Yao, R. et al. DWARF14 is a non-canonical hormone receptor for strigolactone. *Nature* **536**, 469–473 (2016).
6. Shabek, N. et al. Structural plasticity of D3–D14 ubiquitin ligase in strigolactone signalling. *Nature* **563**, 652–656 (2018).
7. Jiang, L. et al. DWARF 53 acts as a repressor of strigolactone signalling in rice. *Nature* **504**, 401–405 (2013); corrigendum **506**, 396 (2014).
8. Zhou, F. et al. D14–$\mathrm{SCF}^{\mathrm{D3}}$-dependent degradation of D53 regulates strigolactone signalling. *Nature* **504**, 406–410 (2013); corrigendum **532**, 402 (2016).
9. Soundappan, I. et al. SMAX1-LIKE/D53 family members enable distinct MAX2-dependent responses to strigolactones and karrikins in *Arabidopsis*. *Plant Cell* **27**, 3143–3159 (2015).
10. Wang, L. et al. Strigolactone signaling in *Arabidopsis* regulates shoot development by targeting D53-like SMXL repressor proteins for ubiquitination and degradation. *Plant Cell* **27**, 3128–3142 (2015).
11. de Saint Germain, A. et al. An histidine covalent receptor and butenolide complex mediates strigolactone perception. *Nat. Chem. Biol*. **12**, 787–794 (2016).
12. Conn, C. E. et al. Convergent evolution of strigolactone perception enabled host detection in parasitic plants. *Science* **349**, 540–543 (2015).
13. Toh, S. et al. Structure–function analysis identifies highly sensitive strigolactone receptors in *Striga*. *Science* **350**, 203–207 (2015).
14. Tsuchiya, Y. et al. Probing strigolactone receptors in *Striga hermonthica* with fluorescence. *Science* **349**, 864–868 (2015).
15. Song, X. et al. IPA1 functions as a downstream transcription factor repressed by D53 in strigolactone signaling in rice. *Cell Res*. **27**, 1128–1141 (2017).
16. Mashiguchi, K. et al. Feedback-regulation of strigolactone biosynthetic genes and strigolactone-regulated genes in *Arabidopsis*. *Biosci. Biotechnol. Biochem*. **73**, 2460–2465 (2009).
17. Braun, N. et al. The pea TCP transcription factor PsBRC1 acts downstream of strigolactones to control shoot branching. *Plant Physiol*. **158**, 225–238 (2012).
18. Scaffidi, A. et al. Strigolactone hormones and their stereoisomers signal through two related receptor proteins to induce different physiological responses in *Arabidopsis*. *Plant Physiol*. **165**, 1221–1232 (2014).
19. Waters, M. T. et al. A *Selaginella moellendorffii* ortholog of KARRIKIN INSENSITIVE2 functions in *Arabidopsis* development but cannot mediate responses to karrikins or strigolactones. *Plant Cell* **27**, 1925–1944 (2015).
20. Lantzouni, O., Klermund, C. & Schwechheimer, C. Largely additive effects of gibberellin and strigolactone on gene expression in *Arabidopsis thaliana* seedlings. *Plant J*. **92**, 924–938 (2017).
21. Duan, J. et al. Strigolactone promotes cytokinin degradation through transcriptional activation of *CYTOKININ OXIDASE/DEHYDROGENASE 9* in rice. *Proc. Natl Acad. Sci. USA* **116**, 14319–14324 (2019).
22. Samodelov, S. L. et al. StrigoQuant: a genetically encoded biosensor for quantifying strigolactone activity and specificity. *Sci. Adv*. **2**, e1601266 (2016).
23. Villaécija-Aguilar, J. A. et al. SMAX1/SMXL2 regulate root and root hair development downstream of KAI2-mediated signalling in *Arabidopsis*. *PLoS Genet*. **15**, e1008327 (2019).
24. Shinohara, N., Taylor, C. & Leyser, O. Strigolactone can promote or inhibit shoot branching by triggering rapid depletion of the auxin efflux protein PIN1 from the plasma membrane. *PLoS Biol*. **11**, e1001474 (2013).
25. Wang, B., Smith, S. M. & Li, J. Genetic regulation of shoot architecture. *Annu. Rev. Plant Biol*. **69**, 437–468 (2018).
26. Kumar, M. N., Hsieh, Y. F. & Verslues, P. E. At14a-Like1 participates in membrane-associated mechanisms promoting growth during drought in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **112**, 10545–10550 (2015).
27. Nelson, D. C. et al. Karrikins enhance light responses during germination and seedling development in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **107**, 7095–7100 (2010).
28. Seale, M., Bennett, T. & Leyser, O. *BRC1* expression regulates bud activation potential but is not necessary or sufficient for bud growth inhibition in *Arabidopsis*. *Development* **144**, 1661–1673 (2017).
29. González-Grandío, E. et al. Abscisic acid signaling is controlled by a BRANCHED1/HD-ZIP I cascade in *Arabidopsis* axillary buds. *Proc. Natl Acad. Sci. USA* **114**, E245–E254 (2017).
30. Liang, Y., Ward, S., Li, P., Bennett, T. & Leyser, O. SMAX1-LIKE7 signals from the nucleus to regulate shoot development in *Arabidopsis* via partially EAR motif-independent mechanisms. *Plant Cell* **28**, 1581–1601 (2016).
31. van Rongen, M., Bennett, T., Ticchiarelli, F. & Leyser, O. Connective auxin transport contributes to strigolactone-mediated shoot branching control independent of the transcription factor *BRC1*. *PLoS Genet*. **15**, e1008023 (2019).
32. Guo, Z. et al. TCP1 modulates brassinosteroid biosynthesis by regulating the expression of the key biosynthetic gene *DWARF4* in *Arabidopsis thaliana*. *Plant Cell* **22**, 1161–1173 (2010).
33. Gonzalez, A., Zhao, M., Leavitt, J. M. & Lloyd, A. M. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J*. **53**, 814–827 (2008).
34. Ling, J. et al. Heat shock protein HSP101 binds to the *Fed-1* internal light regulatory element and mediates its high translational activity. *Plant Cell* **12**, 1213–1227 (2000).
35. Bennett, T. et al. Strigolactone regulates shoot development through a core signalling pathway. *Biol. Open* **5**, 1806–1820 (2016).

Methods

Plant materials and growth conditions

The *smxl6*, *7*, *8* triple mutant[10], *max3-9* mutant[10], *d14-1* mutant[10], *pap1-D* mutant[33] and *35S:SMXL6-GFP* trangenic line[10] have been described previously. T-DNA alleles of *brc1-6* (GK-471F06.17) and the *spl9-4 spl15-1* double mutant (CS67865) were distributed by the Nottingham *Arabidopsis* Stock Centre (NASC). All lines are in the Col-0 background. *Arabidopsis* seeds were surface sterilized, vernalized at 4 °C for 2–4 days, and germinated on half-strength Murashige and Skoog (MS) medium containing 1.0% (w/v) sucrose and 0.7% (w/v) agar. For morphological observations, 10-day-old seedlings were transferred to pots containing a 1:1 vermiculite:soil mixture saturated with one-third-strength MS medium, and grown under a long day photoperiod (16-h light, 8-h dark cycle) with a light intensity of 60–80 $\mu E\, m^{-2}\, s^{-1}$ at 21 °C. For preparation of protoplasts, seedlings were grown under a short-day photoperiod (8-h light, 16-h dark cycle) for 4 weeks as described[10].

Chemicals

Synthetic *rac*-GR24 was obtained as a racemic mixture of equal amounts of GR24^{5DS} and GR24$^{ent\text{-}5DS}$ from Chiralix. The synthetic strigolactone analogues GR24^{4DO}, GR24$^{ent\text{-}4DO}$ and GR24^{5DS} were synthesized as described in the Supplementary Information.

Chemical treatment and RNA-seq analysis

Whole seedlings grown for 10 days horizontally on half-strength MS plates were carefully transferred to half-strength MS liquid medium without injury. Seedlings were cultured in half-strength MS liquid medium for 2 h, then treated with 5 μM GR24^{4DO} or acetone (solvent) in the greenhouse for 2 h or 4 h. Approximately 100-mg whole seedlings were harvested at each time point and frozen in liquid nitrogen as independent biological samples. Total RNAs were isolated using a TRIzol kit (Invitrogen). Illumina sequencing libraries were constructed and sequenced using the Illumina system Hiseq2000 with 125-bp read lengths. The paired-end clean reads of RNA-seq were aligned to the *Arabidopsis* reference genome TAIR10 using STAR (version 2.4.2a)[36]. Fragment quantifications were computed with FeatureCounts (version 1.5.0)[37] in paired-end mode; the features are exons. Default parameters were used for both STAR and FeatureCounts. Expression differentiation analyses were conducted using the R (version 3.3.1) package DESeq (version 1.26.0)[38] with three biological replicates, and genes with fold changes of more than 1.5 and *P* values of less than 0.05 were selected for further analysis. To limit interference by the circadian clock and environmental factors, we compared GR24^{4DO} treatment at 2 h to solvent control at 2 h and to samples at the start point of treatment; genes upregulated in both comparisons were considered to be upregulated strigolactone-responsive genes at 2 h. The same comparison criteria were applied to identify downregulated genes at 2 h and responsive genes at 4 h.

Gene ontology analysis

The enriched functions of DEGs in RNA-seq data sets were annotated with the gene ontology (GO) function[39] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[40] pathways, which were determined using the PlantGSEA program[41].

Gene-expression analysis

Ten-day-old seedlings were cultured in half-strength MS liquid medium for 2 h, then treated with 5 μM GR24^{4DO}, GR24^{5DS}, *rac*-GR24 or acetone (solvent) in the greenhouse for 2 h or 4 h. Approximately 100-mg whole seedlings were harvested at each time point and frozen in liquid nitrogen as independent biological samples. Total RNAs from various materials were extracted using a TRIzol kit (Invitrogen), then treated with TURBO DNase (Ambion). The first-strand complementary DNA was synthesized using oligo (dT) primers with the SSIII first-strand synthesis system (Invitrogen). Real-time PCR experiments were performed using gene-specific primers (Supplementary Table 8) on a CFX 96 real-time PCR detection system (BioRad). *Arabidopsis ACTIN2* (*ACT2*) was used as the internal control. Three or four biological repeats were performed. To compare expression profiles upon chemical treatment, relative expression levels under GR24^{4DO}, GR24^{5DS} or *rac*-GR24 treatment were normalized to expression values of solvent (mock) treatments at specific time points. To compare gene-expression levels in wild-type, mutant and transgenic plants, we normalized the relative expression levels in mutant and transgenic lines to the expression values of the wild type. Gene-expression analysis in buds and 3-week-old plants are described in the figure legends. Experiments were repeated independently three times.

Protein-degradation analysis

35S:SMXL6-GFP transgenic plants were grown for 10 days, collected, and treated with 2 μM *rac*-GR24, GR24^{5DS}, GR24^{4DO} or GR24$^{ent\text{-}4DO}$ for the indicated times in half-strength MS liquid medium at 21 °C. Equal weights of plant materials were collected for protein extraction using lysis buffer (50 mM Tris-HCl at pH 7.5, 150 mM NaCl, 10% glycerol, 1% Nonidet P-40) containing 1× complete protease-inhibitor cocktail (Roche). Protein levels of SMXL6–GFP were detected by immunoblotting with mouse anti-GFP monoclonal antibody (Roche, 11814460001). Experiments were repeated independently three times. For gel source data, see Supplementary Fig. 1.

ChIP analysis

ChIP analyses were performed as described with minor modifications[42]. The *SMXL6:SMXL6-HA 1 s678* transgenic line and *s678* mutant were grown for 2 weeks under the long-day condition on a half-strength MS medium with 1% (w/v) sucrose. Approximately 6-g seedlings were ground into powder in liquid nitrogen and fixed in 1% (v/v) formaldehyde for 30 min at 4 °C to crosslink protein–DNA complexes. The nuclear fraction was isolated and resuspended with ChIP lysis buffer containing 50 mM Tris-HCl, 10 mM EDTA and 1% (v/v) SDS at pH 8.0, then placed on ice for 30 min. Two volumes of ChIP dilution buffer containing 16.7 mM Tris-HCl, 167 mM NaCl and 1.1% (v/v) Triton X-100 at pH 8.0 were added to the samples before sonication for 14 min (30 s on, 30 s off, high level) in a Bioruptor (Diagenode, UCD-200) to yield DNA fragments of 300–500 base pairs in length. The lysates were diluted using seven volumes of ChIP dilution buffer and centrifuged at 16,000*g* for 15 min at 4 °C. Approximately 1.25% (v/v) of the sample was kept as input, and the remainder of the supernatant was mixed with an anti-HA polyclonal antibodies (Sigma, H6908) coupled to Dynabeads protein G (Life Technologies, 10003D), and incubated overnight at 4 °C. The protein–DNA complexes were washed and eluted, and then underwent reverse crosslinking. The DNA was precipitated in ethanol, recovered, dissolved in water and stored at −80 °C.

Analysis of ChIP–seq data

ChIP–seq libraries were prepared using the NextflexTM Rapid DNA-seq Kit (Bioo Scientific, 5144-02) and sequenced using the Illumina system Novaseq 6000 with a 150-bp read length. All ChIP–seq reads were mapped to the *Arabidopsis* genome TAIR10 using BWA (version 0.7.10–r789) software[43] with default parameters. Duplicated reads and reads with low mapping quality were discarded using SAMtools (version 0.1.19–44428cd)[44] with default parameters. Peaks in the *SMXL6:SMXL6-HA 1 s678* transgenic line were identified by comparison with the *s678* mutant using the model-based analysis software MACS (version 1.4.2)[45], with parameters '–mfold = 10, 30', then determined by *P* values of less than 1×10^{-5} and fold changes of more than 2.0. All detected peaks are listed in Supplementary Table 4 and were used for further analysis.

ChIP–qPCR

The prepared DNA in ChIP was applied in SsoFast EvaGreen supermix (BioRad) with a BioRad CFX96 real-time PCR detection system.

Chromatin of *s678* seedlings was used as a control. The enrichment of target gene promoters is displayed as a percentage of the input DNA. The *TUA2* promoter was used as a nonspecific target. All primers used in qPCR analyses are listed in Supplementary Table 8. Experiments were repeated independently three times.

Recombinant protein purification and EMSAs

The full-length coding sequences of *SMXL6*, *7*, *8* were amplified and cloned into the expression vector pGEX 6p-1 (GE Healthcare). The GST, GST–SMXL6, GST–SMXL7 and GST–SMXL8 proteins were induced using 0.3 mM isopropyl β-D-1-thiogalactopyranoside at 16 °C for 20 h in BL21 transetta cells (Transgene). Fusion proteins were purified using glutathione sepharose 4 fast flow (GE Healthcare) and quantified by SDS–PAGE. Biotin 5′-end-labelled DNA probes were synthesized or amplified using biotin 5′-end-labelled primers. Binding reaction samples containing 0.2 ng or 10 fmol of biotin-labelled double-stranded DNAs, 0.5 μg of the recombinant GST, GST–SMXL6, GST–SMXL7 or GST–SMXL8 protein and 1 μg poly(dIdC) were incubated at room temperature for 40 min and subjected to electrophoresis on 4% (w/v) polyacrylamide gels with half-strength TBE buffer. Biotin-labelled DNA was detected using the Lightshift chemiluminescent EMSA kit (Thermo Scientific). The supershift band was detected in the presence of 3.7 μg anti-GST monoclonal antibody (Sigma, G1160). All primers and probes are listed in Supplementary Table 8. Plasmid sequences have been deposited in GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under the accession numbers listed in Supplementary Table 9. Experiments were repeated independently three or four times. For gel source data, see Supplementary Fig. 1.

Leaf morphology analysis

The fifth leaves of 3-week-old plants were harvested, laid flat on the surface of an agar medium plate, and photographed for further analysis. The leaf length (distance between the leaf tip and the base of petiole) and leaf width (the greatest distance across the leaf lamina perpendicular to the proximal/distal axis of the leaf) were measured manually using ImageJ (1.41o) software (http://rsbweb.nih.gov/ij/) as described[10].

Anthocyanin measurement

Three-week-old *Arabidopsis* seedlings were collected, ground into powder in liquid nitrogen, quickly weighed and boiled for 3 min in extraction buffer (propanol:HCl:H_2O = 18:1:81). The anthocyanin content is presented as (A535 − A650) per gram of fresh weight[46].

Quantification of ABA content

For determination of the ABA content, unelongated buds in the axils of rosette leafs were carefully collected, frozen and ground in liquid nitrogen. Approximately 30 mg of frozen samples were weighed accurately, and extracted with 1 ml methanol containing 0.5 ng D_6-ABA (internal standard) overnight at −20 °C. Samples were centrifuged for 15 min at 15,000*g* and 4 °C, and the supernatant was collected, evaporated under nitrogen gas, and dissolved in ammonia solution (5% v/v). Then ABA was purified using Oasis MAX (Waters) solid-phase extraction cartridges. Endogenous ABA was analysed as described[47] with some modifications. Liquid chromatography with tandem mass spectrometry analysis was performed on an ultraperformance liquid chromatography (UPLC) system (Waters) coupled to a 6500 Q-Trap system (AB SCIEX). Liquid-chromatography separation used a BEH C_{18} column (internal diameter 2.1 mm × 100 mm, 1.7 μm; Waters) with mobile phases A (0.05% v/v acetic acid in water) and B (0.05% v/v acetic acid in acetonitrile). The gradient was set with an initial 20% of mobile phase B, increased to 70% B within 6 min. ABA was detected in multiple reaction monitoring (MRM) mode with transition 263.0/153.1. Quantitation was performed using the isotope dilution method.

Drought stress

Plants were grown on half-strength MS plates for a week, transferred to pots containing a 1:1 vermiculite:soil mixture saturated with one-third-strength MS medium, and grown in greenhouses (16-h light/8-h dark cycle, 60–80 μE m^{-2} s^{-1} light intensity, 21–22 °C) for a week before exposure to drought stress. Drought stress was carried out by withholding water until the lethal effect of dehydration was observed. The survival rate was calculated after rewatering for 3 days.

Transcriptional activity assay in protoplasts

To construct the recombinant plasmids used in transcriptional activity assays, we amplified the promoters of *SMXL6*, *SMXL7*, *SMXL8* and *BRC1* and inserted them into the 35SLUC vector using an In-Fusion PCR cloning kit (Clontech), with primers listed in Supplementary Table 8. To detect regulation by SMXL6 and SMXL6$^{\Delta EAR}$ on the promoters of *SMXL6*, *SMXL7*, *SMXL8* and *BRC1*, we introduced combinations of effector plasmids (35S-GFP, 3×Flag-SMXL6 or 3×Flag-SMXL6-no-EAR), reporter plasmids (pSMXL6-LUC, pSMXL7-LUC, pSMXL8-LUC or pBRC1-LUC) and reference plasmids (pRTL) into *Arabidopsis* protoplasts as described[10]. To examine the influence of GR24^{4DO} on SMXL6-mediated repression at the promoters of *SMXL6*, *7*, *8* and *BRC1*, we introduced combinations of effector plasmids (35S-GFP or 3×Flag-SMXL6), reporter plasmids (pSMXL6-LUC, pSMXL7-LUC, pSMXL8-LUC or pBRC1-LUC) and reference plasmids (pRTL) into *Arabidopsis* protoplasts. After incubation in the prescence or absence of 50 μM GR24^{4DO} at 21 °C for 16 h, luciferase activities were measured using a dual-luciferase reporter assay system (Promega). To detect SMXL6-mediated regulation of the *SMXL7-WT* and *SMXL7-mu3* promoters, we introduced combinations of effector plasmids (35S-GFP or 3×Flag-SMXL6) and reporter plasmids (pSMXL7-WT-LUC or pSMXL7-mu3-LUC) into *Arabidopsis* protoplasts. After incubation at 21 °C for 12 h, luciferase activities were measured using the dual-luciferase reporter assay system (Promega). Plasmid sequences have been deposited in GenBank under the accession numbers listed in Supplementary Table 9. Experiments were repeated independently three times.

Plasmid construction and plant transformation

To construct the pSMXL6-SMXL6-HA and pSMXL6-SMXL6-no-EAR-HA plasmids, we amplified the 3×HA sequence, the nopaline synthase (NOS) terminator sequence, and the promoter and coding sequences of *SMXL6* and *SMXL6$^{\Delta EAR}$* and sequentially cloned them into the pCAMBIA1300 vector. To construct p35S-TCP1-SRDX, we amplified the coding sequence of *TCP1* and cloned it into the pWM101 vector. For CRISPR analysis of *PAP1* and *PAP2*, the designed PAPs-targeting sequences were cloned into the AtU6-26-target-sgRNA vector, and then into the pYAO:hSpCas9 vector[48]. For CRISPR analysis of *TCP1*, the designed *TCP1*-targeting sequences were cloned into the pHEE401 vector[49]. All recombinant plasmids were introduced into *Agrobacterium tumefaciens* strain EHA105 and transformed into indicated *Arabidopsis* recipients as reported previously[50]. The primers are listed in Supplementary Table 8. GenBank accession numbers of plasmid sequences are listed in Supplementary Table 9.

Genetic analysis

The double, triple and quadruple mutants were generated by crossing relevant homozygous single or double mutants and genotyping the F_2 or F_3 progenies. The genotypes of *smxl6*, *smxl7*, *smxl8* and *max3-9* mutants were detected as described[10]. The *brc1-6* mutant was genotyped using primer pairs of *brc1-6* RP plus GABI-Kat and *brc1-6* LP plus *brc1-6* RP. The *spl9-4* mutant was genotyped using primer pairs of *spl9-4* RP plus LB1 and *spl9-4* LP plus *spl9-4* RP. The *spl15-1* mutant was genotyped using primer pairs of *spl15-1* RP plus LBb1 and *spl15-1* LP plus *spl15-1* RP. The primers are listed in Supplementary Table 8.

Accession numbers

Sequence data from this article can be found in GenBank/EMBL (https://www.ncbi.nlm.nih.gov/genbank/) under the following accession numbers: *ABCG37* (AT3G53480), *ACT2* (AT3G18780), *AFL1* (AT3G28270),

ANS (AT4G22880), *BEE3* (AT1G73830), *BGLU47* (AT4G21760), *bHLH29* (AT2G28160), *BRC1* (AT3G18550), *DFR* (AT5G42800), *DLK2* (AT3G24420), *DVL15* (AT3G46613), *ELIP1* (AT3G22840), *EXP11* (AT1G20190), *EXP12* (AT3G15370), *HB40* (AT4G36740), *HB52* (AT5G53980), *HEC1* (AT5G67060), *HSFA6B* (AT3G22830), *IAA19* (AT3G15540), *IAA29* (AT4G32280), *KAN1* (AT5G16560), *KUF1* (AT1G31350), *MAX3* (AT2G44990), *MAX4* (AT4G32810), *MYB113* (AT1G66370), *MYB114* (AT1G66380), *MYB27* (AT3G53200), *NRT2.6* (AT3G45060), *PAP1* (AT1G56650), *PAP2* (AT1G66390), *PAR1* (AT2G42870), *REM36* (AT4G31620), *RR16* (AT2G40670), *SAUR3* (AT4G34790), *SAUR22* (AT5G18050), *SAUR29* (AT3G03820), *SAUR61* (AT1G29420), *SAUR65* (AT1G29460), *SMXL6* (AT1G07200), *SMXL7* (AT2G29970), *SMXL8* (AT2G40130), *SPL9* (AT2G42200), *SPL15* (AT3G57920), *STH7* (AT4G39070), *TCP1* (AT1G67260), *TGA7* (AT1G77920), *TT7* (AT5G07990), *TUA2* (AT1G50010), *WRKY38* (AT5G22570), *WRKY49* (AT5G43290), *YUC3* (AT1G04610), AT1G64380, AT1G71520 and AT5G56840.

Statistical analysis

For gene-expression analyses, observations of phenotype, measurement of ABA and anthocyanins, ChIP–qPCR and assays of transcriptional activity, statistical analysis was assessed as described in the figure legends. *P* values were calculated by two-sided Student's *t*-tests using Excel 2016, or by Tukey's HSD test using R (version 3.6.1), and are shown in bar graphs. Statistical analyses of RNA-seq and ChIP–seq data are described in the Methods sections 'Chemical treatment and RNA-seq analysis' and 'Analysis of ChIP–seq data'. We used Fisher's exact test to calculate the enrichment of GR24^{4DO}-responsive genes amongst SMXL6–HA-targeted genes, and the χ^2 test to calculate the enrichment of ATAACAA in flanking sequences around the ChIP–seq peak summits of GR24^{4DO}-responsive genes. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Uncropped gels and blots are provided in Supplementary Fig. 1. The RNA-seq and ChIP–seq data have been deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under the accession numbers GSE126331 and GSE140705. Materials and reagents are available from the corresponding authors on request. Source data are provided with this paper.

36. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
37. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
38. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **11**, R106 (2010).
39. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet*. **25**, 25–29 (2000).
40. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. **40**, D109–D114 (2012).
41. Yi, X., Du, Z. & Su, Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res*. **41**, W98–W103 (2013).
42. Cui, X. et al. REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in *Arabidopsis*. *Nat. Genet*. **48**, 694–699 (2016).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. **9**, R137 (2008).
46. Song, S. et al. The bHLH subgroup IIId factors negatively regulate jasmonate-mediated plant defense and development. *PLoS Genet*. **9**, e1003653 (2013).
47. Fu, J., Chu, J., Sun, X., Wang, J. & Yan, C. Simple, rapid, and simultaneous assay of multiple carboxyl containing phytohormones in wounded tomatoes by UPLC-MS/MS using single SPE purification and isotope dilution. *Anal. Sci*. **28**, 1081–1087 (2012).
48. Yan, L. et al. High efficiency genome editing in *Arabidopsis* using *YAO* promoter-driven CRISPR/Cas9 system. *Mol. Plant* **8**, 1820–1823 (2015).
49. Wang, Z. P. et al. Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in *Arabidopsis* in a single generation. *Genome Biol*. **16**, 144 (2015).
50. Mou, Z. et al. Silencing of phosphoethanolamine *N*-methyltransferase results in temperature-sensitive male sterility and salt hypersensitivity in Arabidopsis. *Plant Cell* **14**, 2031–2043 (2002).

Acknowledgements We thank J. Qiu for providing the *pap1-D* mutant; Q. Xie for providing the CRISPR–CAS9 system driven by the *YAO* promoter; and Q. Chen for providing the CRISPR–CAS9 system driven by the egg-cell-specific promoter. This work was supported by grants from the National Natural Science Foundation of China (31788103, 31600221 and 31661143025) and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2019099).

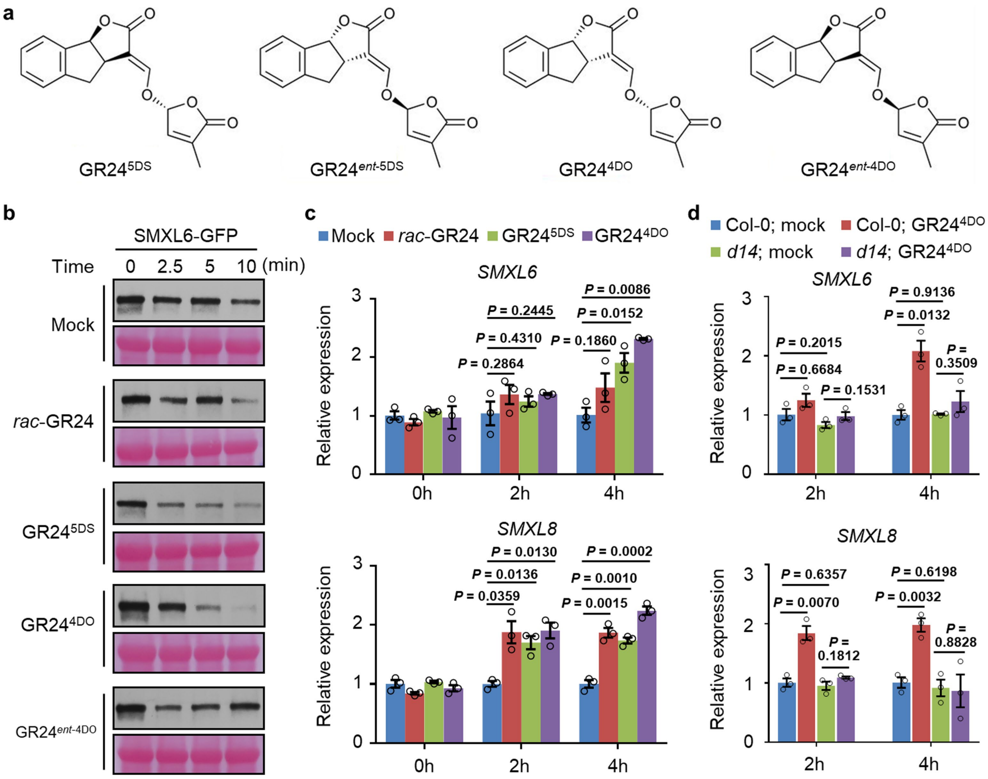
Author contributions J.L. and B.W. designed and supervised the overall research. L.W. and B.W. carried out chemical treatments, gene-expression analysis and phenotype observations. L.W. carried out protein-degradation analysis, ChIP analysis, EMSAs, anthocyanin measurements and assays of transcriptional activity. H.Y., T.L. and L.K. processed RNA-seq and ChIP–seq data. H.G., A.W., N.S. and H.M. contributed to genetic analyses and plasmid construction. X.L. and J.Y. synthesized the strigolactone analogues GR24^{4DO}, GR24$^{ent\text{-}4DO}$ and GR24^{5DS}. J.C. and B.W. measured ABA content. L.W., B.W. and G.X. performed statistical analyses. B.W., L.W. and J.L. analysed data and wrote the paper with input from all other authors.

Competing interests The authors declare no competing interests.

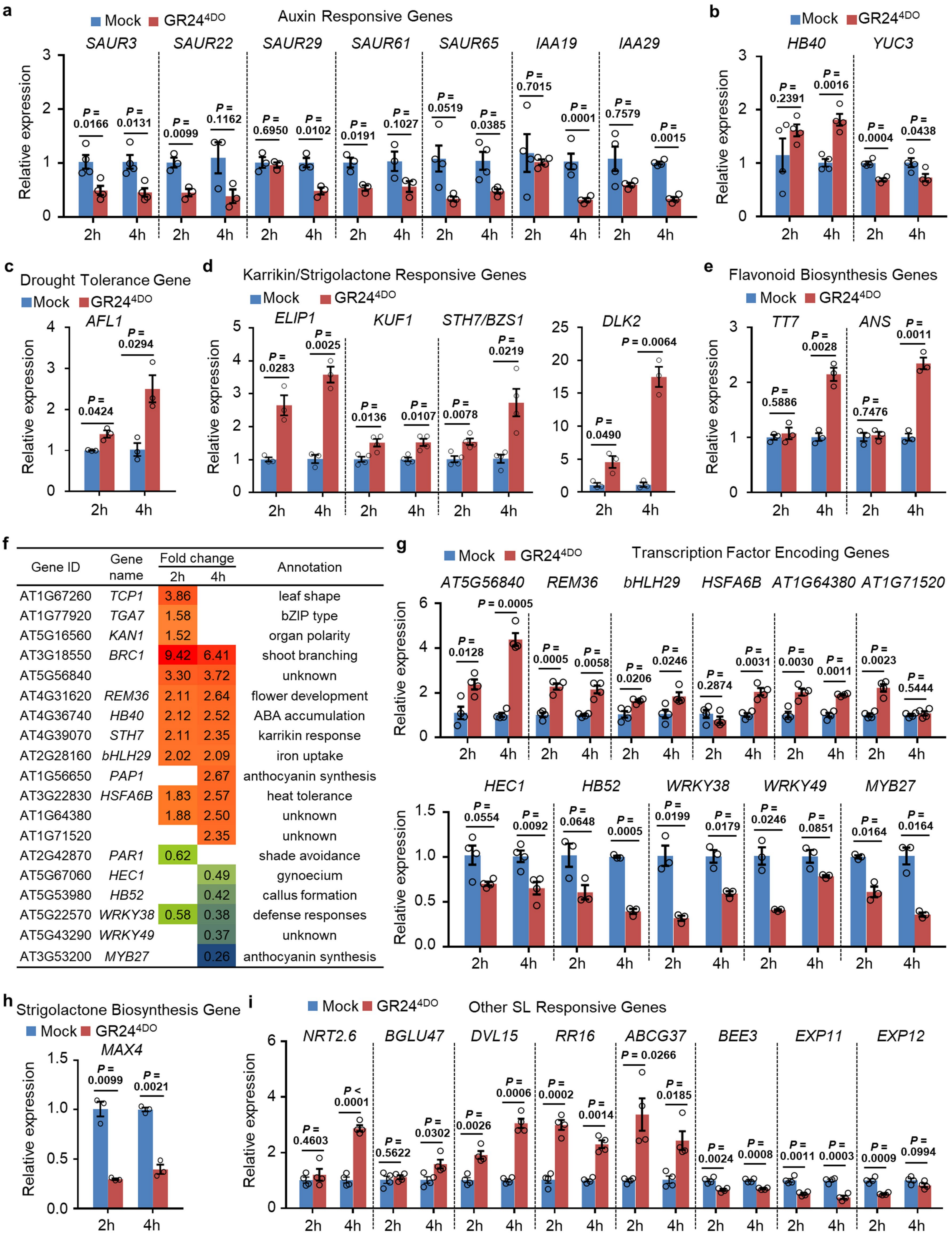
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2382-x.

Correspondence and requests for materials should be addressed to B.W. or J.L.



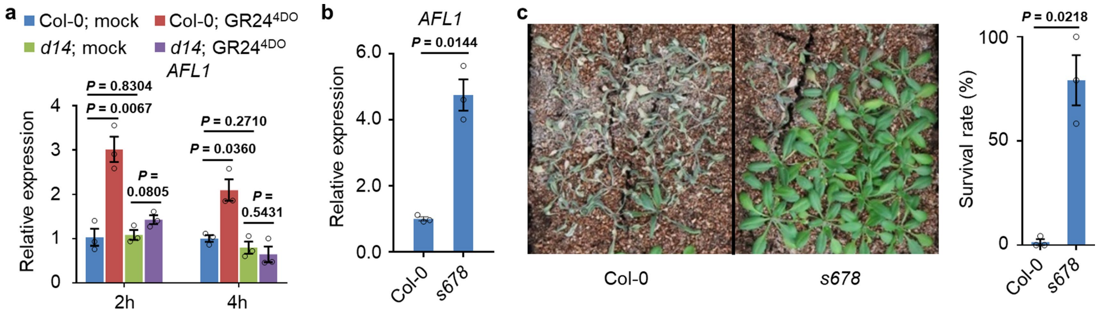
Extended Data Fig. 1 | Comparison of *rac*-GR24 and GR24 stereoisomers in stimulating strigolactone signalling. **a**, Chemical structures of GR24 stereoisomers with different stereochemical features. **b**, Stability of SMXL6–GFP in *35S:SMXL6-GFP* transgenic seedlings treated with 2 μM *rac*-GR24, GR24^{5DS}, GR24^{4DO} or GR24$^{ent\text{-}4DO}$. Proteins were detected by immunoblotting with an anti-GFP monoclonal antibody. Data represent three independent experiments. **c**, Expression of *SMXL6* and *SMXL8* in 10-day-old wild-type seedlings pretreated in half-strength MS liquid medium for 2 h, and then treated with 5 μM *rac*-GR24, GR24^{5DS} or GR24^{4DO} for 0 h, 2 h or 4 h. Data were normalized to mock treatment at specific time points and are means ± s.e.m.; *n* = 3 biologically independent samples. **d**, Induction of *SMXL6* and *SMXL8* upon GR24^{4DO} treatment is blocked by *D14* mutation. Data were normalized to mock treatment in wild-type seedlings at specific time points and are means ± s.e.m.; *n* = 3 biologically independent samples. **c**, **d**, *P* values are shown; two-sided Student's *t*-test.



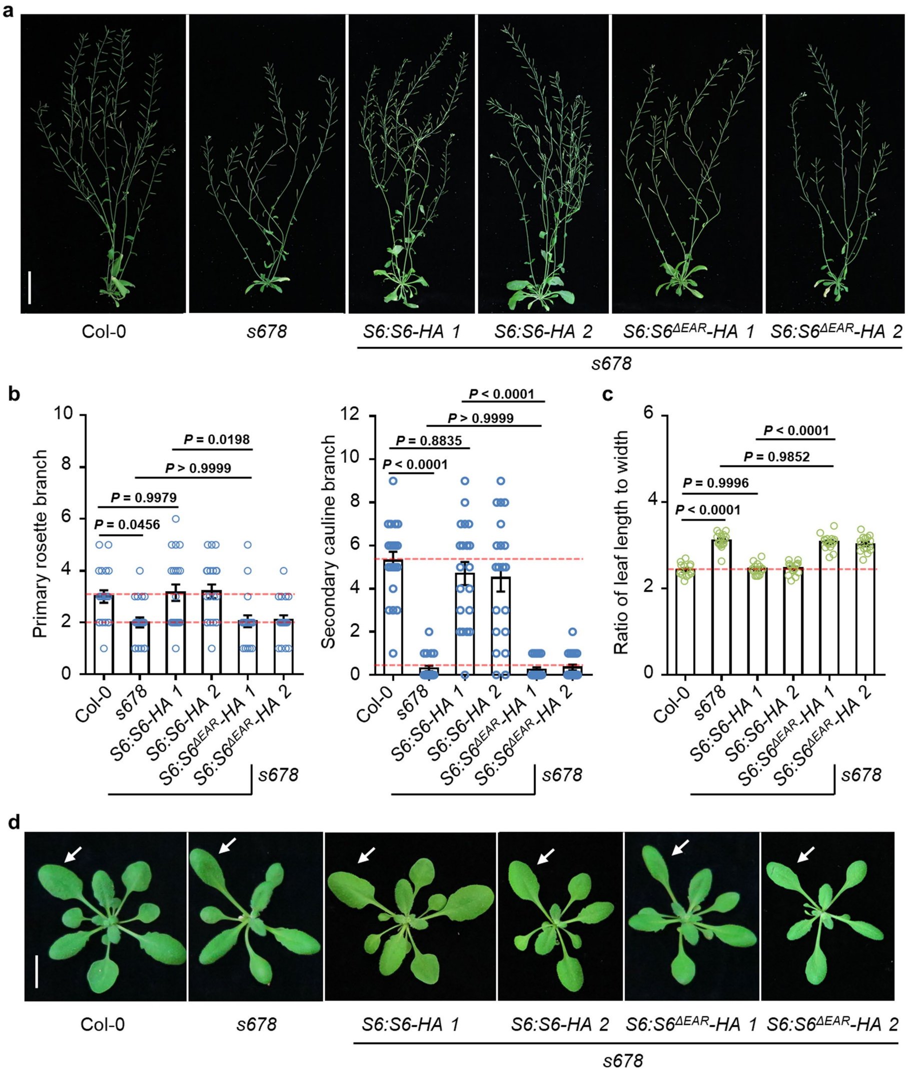
Gene ID	Gene name	Fold change 2h	Fold change 4h	Annotation
AT1G67260	*TCP1*	3.86		leaf shape
AT1G77920	*TGA7*	1.58		bZIP type
AT5G16560	*KAN1*	1.52		organ polarity
AT3G18550	*BRC1*	9.42	6.41	shoot branching
AT5G56840		3.30	3.72	unknown
AT4G31620	*REM36*	2.11	2.64	flower development
AT4G36740	*HB40*	2.12	2.52	ABA accumulation
AT4G39070	*STH7*	2.11	2.35	karrikin response
AT2G28160	*bHLH29*	2.02	2.09	iron uptake
AT1G56650	*PAP1*		2.67	anthocyanin synthesis
AT3G22830	*HSFA6B*	1.83	2.57	heat tolerance
AT1G64380		1.88	2.50	unknown
AT1G71520			2.35	unknown
AT2G42870	*PAR1*	0.62		shade avoidance
AT5G67060	*HEC1*		0.49	gynoecium
AT5G53980	*HB52*		0.42	callus formation
AT5G22570	*WRKY38*	0.58	0.38	defense responses
AT5G43290	*WRKY49*		0.37	unknown
AT3G53200	*MYB27*		0.26	anthocyanin synthesis

Extended Data Fig. 2 | See next page for caption.

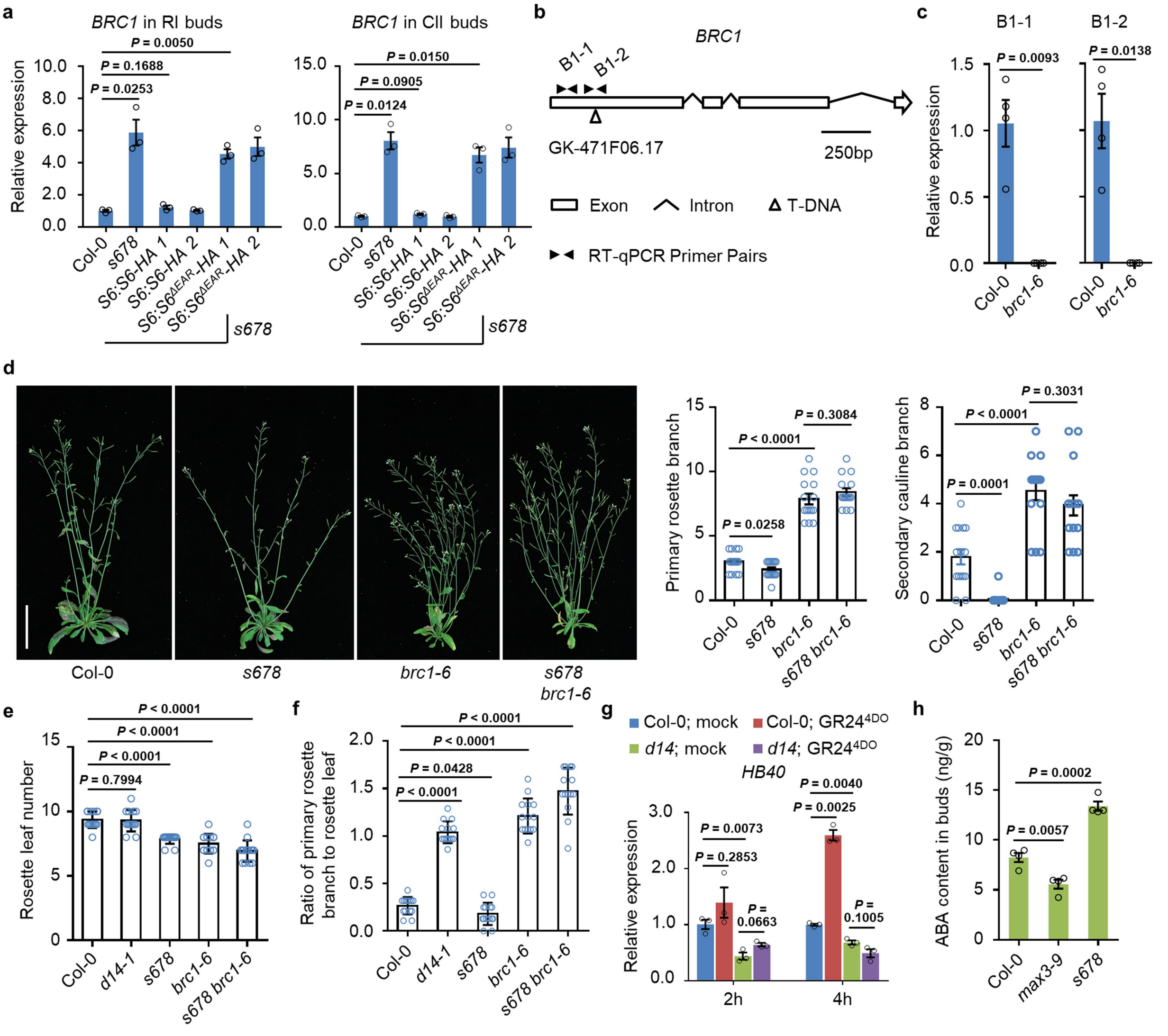
Extended Data Fig. 2 | Verification of strigolactone-responsive genes by qRT–PCR analyses. a–**e**, **g**–**i**, Expression of strigolactone-responsive genes in 10-day-old wild-type (Col-0) seedlings pretreated in half-strength MS liquid medium for 2 h, and then treated with 5 μM GR24^{4DO} in half-strength MS liquid medium for 2 h or 4 h. The qRT–PCR data in **a**–**e**, **g**–**i** are means ± s.e.m.; $n = 3$ biologically independent samples (for *SAUR22, SAUR29, SAUR61, AFL1, ELIP1, DLK2, TT7, ANS, HB52, WRKY38, WRKY49, MYB27, MAX4*) or 4 biologically independent samples (for *SAUR3, SAUR65, IAA19, IAA29, HB40, YUC3, KUF1, STH7/BZS1, AT5G56840, REM36, bHLH29, HSFA6B, AT1G64380, AT1G71520, HEC1, NRT2.6, BGLU47, DVL15, RR16, ABCG37, BEE3, EXP11, EXP12*). *P* values are shown; two-sided Student's *t*-test. **f**, Fold change in transcription-factor-encoding genes in response to GR24^{4DO} treatment based on RNA-seq data. *ABCG37, ATP-BINDING CASSETTE G37*; *AFL1, At14a-LIKE 1*; *ANS, ANTHOCYANIDIN SYNTHASE*; *BEE3, BR-ENHANCED EXPRESSION 3*; *BGLU47, BETA-GLUCOSIDASE 47*; *bHLH29, BASIC HELIX-LOOP-HELIX PROTEIN 29*; *BRC1, BRANCHED 1*; *DLK2, D14-LIKE 2*; *DVL15, DEVIL 15*; *ELIP1, EARLY LIGHT-INDUCIBLE PROTEIN 1*; *EXP11, EXP12, EXPANSIN 11, 12*; *HB40, HB52, HOMEOBOX PROTEIN 40, 52*; *HEC1, HECATE 1*; *HSFA6B, HEAT SHOCK TRANSCRIPTION FACTOR A6B*; *IAA19, IAA29, INDOLE-3-ACETIC ACID INDUCIBLE 19, 29*; *KAN1, KANADI 1*; *KUF1, KAR-UP F-BOX 1*; *MAX4, MORE AXILLARY GROWTH 4*; *MYB27, MYB DOMAIN PROTEIN 27*; *NRT2.6, NITRATE TRANSPORTER 2.6*; *PAP1, PRODUCTION OF ANTHOCYANIN PIGMENT 1*; *PAR1, PHY RAPIDLY REGULATED 1*; *REM36, REPRODUCTIVE MERISTEM 36*; *RR16, RESPONSE REGULATOR 16*; *SAUR3, SAUR22, SAUR29, SAUR61, SAUR65, SMALL AUXIN UPREGULATED RNA 3, 22, 29, 61, 65*; *STH7/BZS1, SALT TOLERANCE HOMOLOG 7/BZR1-1D SUPPRESSOR 1*; *TCP1, TCP DOMAIN PROTEIN 1*; *TGA7, TGACG SEQUENCE-SPECIFIC BINDING PROTEIN 7*; *TT7, TRANSPARENT TESTA 7*; *WRKY38, WRKY49, WRKY DNA-BINDING PROTEIN 38, 49*; *YUC3, YUCCA 3*.



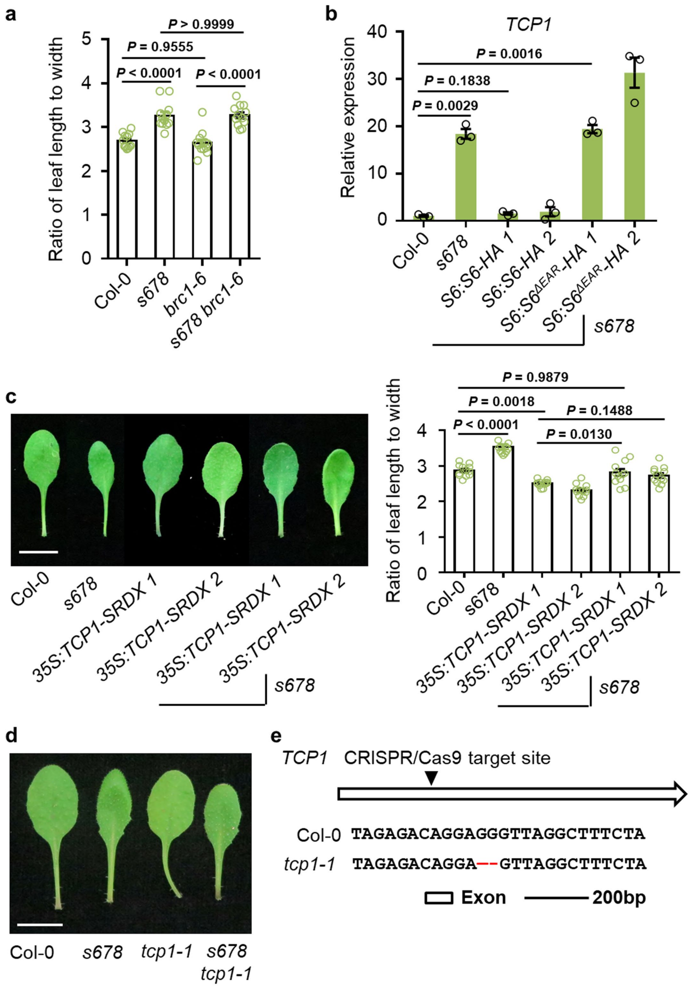
Extended Data Fig. 3 | SMXL6, 7, 8 negatively regulate *AFL1* expression and drought tolerance. a, Induction of *AFL1* upon GR24^{4DO} treatment is blocked in the *d14* mutant. Data were normalized to mock treatment in wild-type seedlings at specific time points and are means ± s.e.m.; $n = 3$ biologically independent samples. **b**, Expression of *AFL1* in 10-day-old seedlings of the wild type (Col-0) and *s678* triple mutant. Data were normalized to the wild type and are means ± s.e.m.; $n = 3$ biologically independent samples. **c**, Phenotype and survival rate of wild-type and *s678* plants. We exposed 2-week-old plants to drought stress by withholding of water for 2 weeks and then rewatered for 3 days. Data are means ± s.e.m.; $n = 3$ pools (24 plants each pool). **a**–**c**, *P* values are shown; two-sided Student's *t*-test.



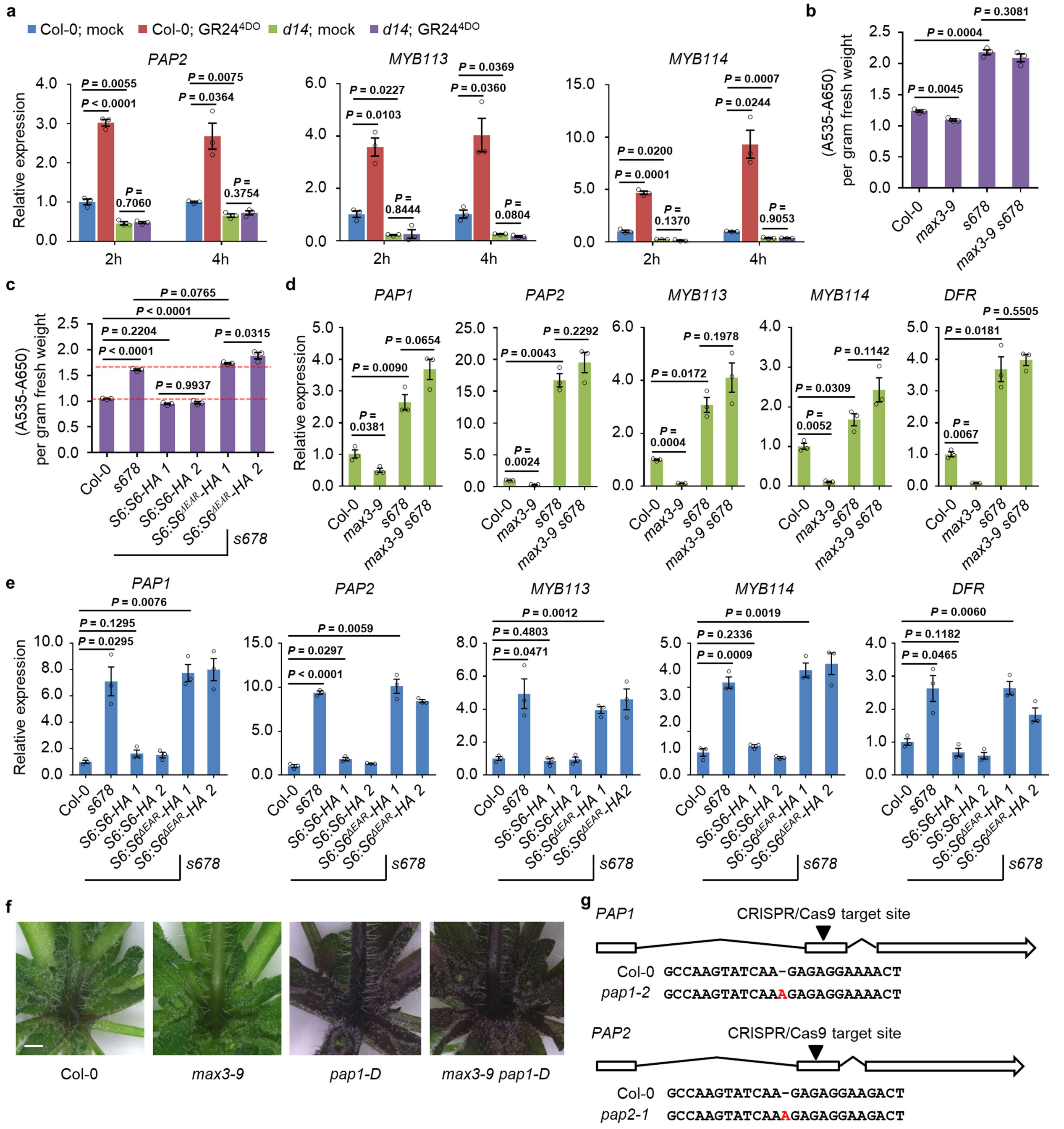
Extended Data Fig. 4 | The EAR motif of SMXL6 is essential for regulation of shoot and leaf development. **a**, Phenotypes of wild-type (Col-0), *s678*, *S6:S6-HA s678* and *S6:S6*$^{\Delta EAR}$*-HA s678* plants. *S6:S6-HA 1 s678* and *S6:S6-HA 2 s678* represent independent transgenic lines with the *S6:S6-HA* transgene in the *s678* background; *S6:S6*$^{\Delta EAR}$*-HA 1 s678* and *S6:S6*$^{\Delta EAR}$*-HA 2 s678* represent independent transgenic lines containing the *S6:S6*$^{\Delta EAR}$*-HA* transgene in the *s678* background. Scale bar, 5 cm. **b**, Quantitative analysis of shoot branching in the adult plants shown in **a**. We counted the number of primary branches grown from the rosette leaf axil (left) and secondary branches grown from the cauline leaf axil (right) of at least 0.5 cm. Data are means ± s.e.m.; $n = 20$ plants. **c**, Ratio of leaf length to width for the fifth leaves of the wild-type (Col-0), *s678*, *S6:S6-HA s678* and *S6:S6*$^{\Delta EAR}$*-HA s678* plants after growth for 3 weeks. Data are means ± s.e.m.; $n = 15$ leaves. **d**, Leaf morphology of 3-week-old plants. The fifth leaves are marked by white arrows. Scale bar, 1 cm. **b**, **c**, P values are shown; Tukey's HSD test.



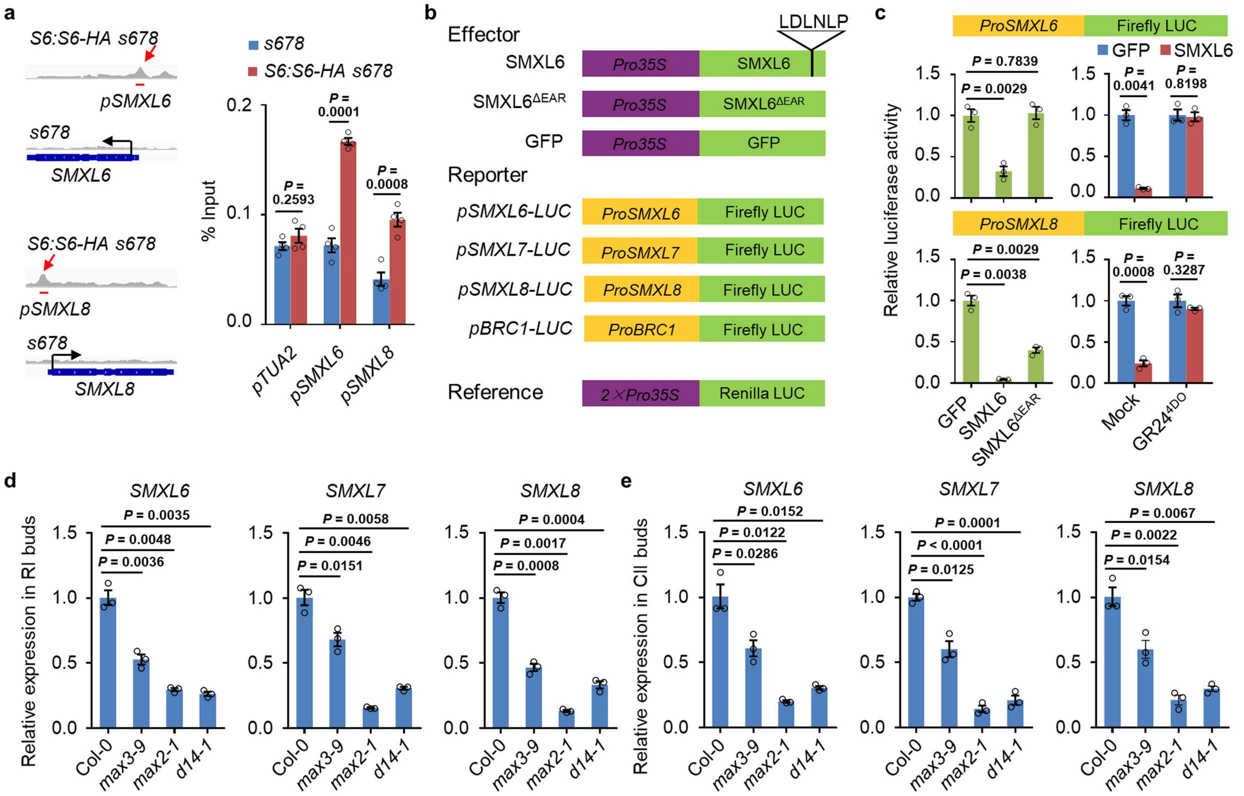
Extended Data Fig. 5 | *BRC1* regulates strigolactone-mediated shoot branching. **a**, Expression of *BRC1* in buds of primary rosette (RI) and secondary cauline (CII) branches. Data were normalized to the wild type (Col-0) and are means ± s.e.m.; *n* = 3 biologically independent samples. **b**, Schematic representation of the T-DNA insertion mutant *brc1-6*. The T-DNA insertion site and positions of RT–qPCR primers are indicated by white and black triangles, respectively. **c**, Expression of *BRC1* in 10-day-old seedlings using B1-1 and B1-2 primer pairs. Data were normalized to the wild type and are means ± s.e.m.; *n* = 3 biologically independent samples. **d**, Shoot-branching phenotypes of plants at the adult stage. Scale bar, 5 cm. Data represent 15 independent experiments. We counted the number of primary rosette branches and secondary cauline branches of at least 0.5 cm. Data are means ± s.e.m.; *n* = 15 plants. **e**, **f**, Rosette leaf number (**e**) and the ratio of primary rosette branch to rosette leaf number (**f**) in adult Col-0, *d14-1*, *s678*, *brc1-6*, and *s678 brc1-6* plants. Data are means ± s.e.m.; *n* = 14 plants. **g**, Induction of *HB40* upon GR24^{4DO} treatment is blocked in the *d14* mutant. Data were normalized to mock treatment in wild-type seedlings at specific time points and are means ± s.e.m.; *n* = 3 biologically independent samples. **h**, ABA content in unelongated axillary buds of rosette leaves in Col-0, *max3-9* and *s678* plants. Data are means ± s.e.m.; *n* = 4 biologically independent samples. **a**, **c**–**h**, *P* values are shown; two-sided Student's *t*-test.



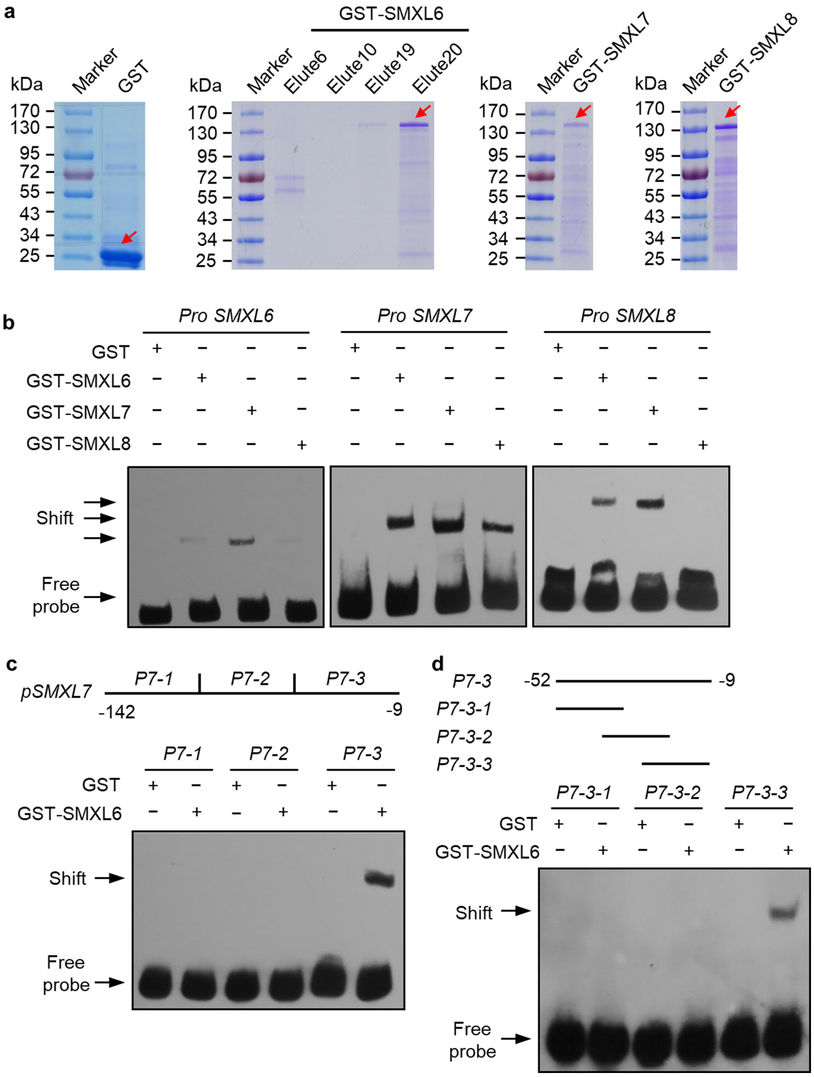
Extended Data Fig. 6 | *TCP1* contributes to strigolactone-mediated leaf-shape regulation. **a**, Ratio of leaf length to width for the fifth leaves in wild-type (Col-0), *s678*, *brc1-6* and *s678 brc1-6* plants grown for 3 weeks. Data are means ± s.e.m.; *n* = 12 leaves. **b**, The EAR motif is required for SMXL6-mediated regulation of the expression of *TCP1* in seedlings. Data were normalized to wild type and are means ± s.e.m.; *n* = 3 biologically independent samples. **c**, Left, morphology of the fifth leaves of 3-week-old plants. *35S:TCP1-SRDX 1* and *35S:TCP1-SRDX 2* represent independent transgenic lines with the *35S:TCP1-SRDX* transgene in the Col-0 background; *35S:TCP1-SRDX 1 s678* and *35S:TCP1-SRDX 2 s678* represent independent transgenic lines with the *35S:TCP1-SRDX* transgene in the *s678* background. Scale bar, 1 cm. Right, ratio of leaf length to width for the fifth leaves. Data are means ± s.e.m.; *n* = 12 leaves. **d**, Morphology of the fifth leaves of wild-type, *s678*, *tcp1-1* and *s678 tcp1-1* plants grown for 3 weeks. Scale bar, 1 cm. **e**, Mutation site of the *tcp1-1* mutant. **a**–**c**, *P* values are shown; Tukey's HSD test (**a**, **c**) or two-sided Student's *t*-test (**b**).



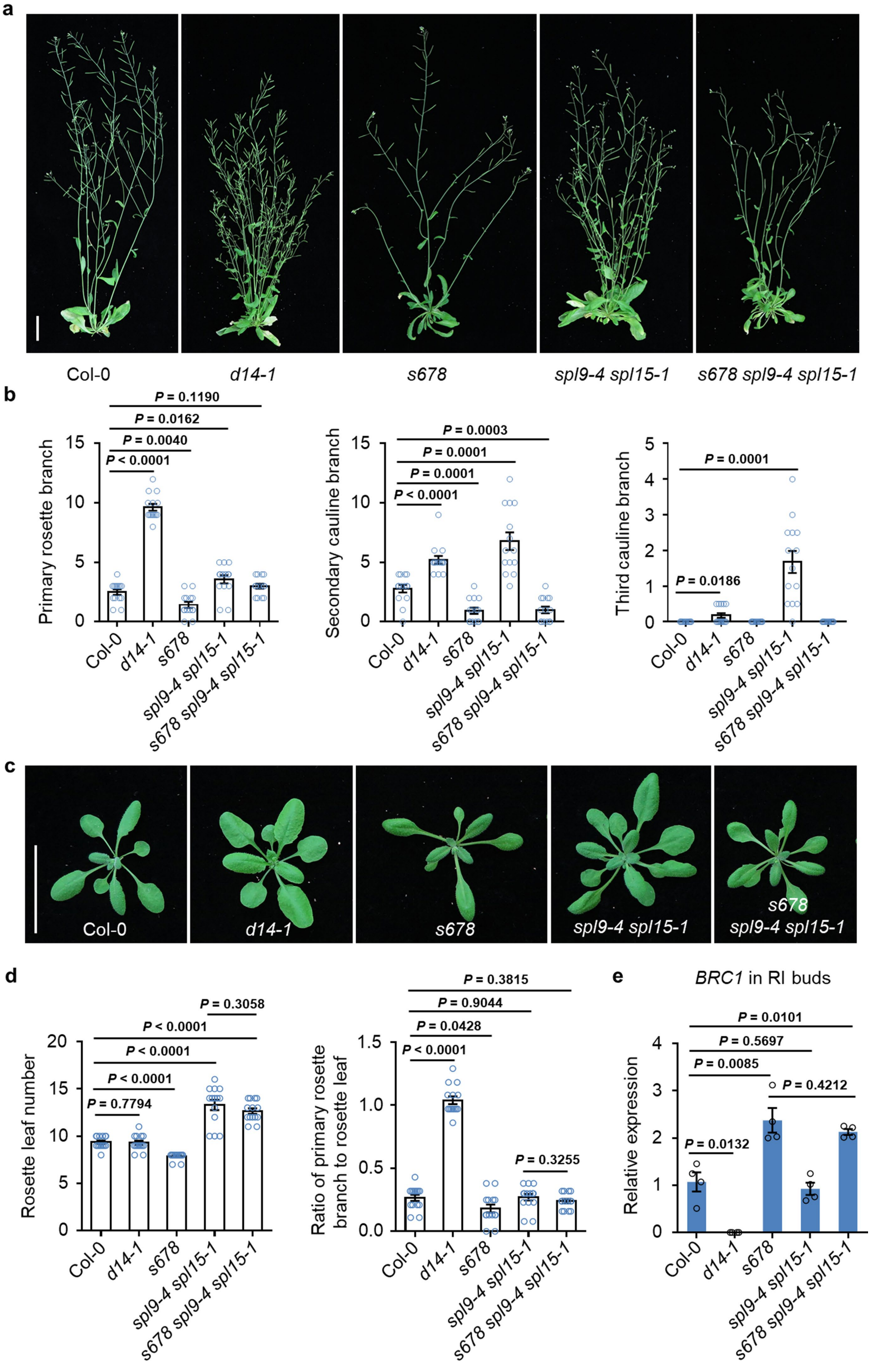
Extended Data Fig. 7 | Strigolactones promote expression of anthocyanin-biosynthesis genes. a, Induction of *PAP2*, *MYB113* and *MYB114* upon GR24^{4DO} treatment is blocked in the *d14* mutant. Data were normalized to mock treatment in wild-type seedlings at specific time points and are means ± s.e.m.; *n* = 3 biologically independent samples. **b**, **c**, Anthocyanin content in Col-0 plants, strigolactone mutants and transgenic lines after growth for 3 weeks. Data are means ± s.e.m.; *n* = 3 pools (6 seedlings per pool). **d**, Expression of *PAP1*, *PAP2*, *MYB113*, *MYB114* and *DFR* in 3-week-old wild-type (Col-0), *max3-9*, *s678* and *max3-9 s678* plants. Data were normalized to wild type and are means ± s.e.m.; *n* = 3 biologically independent samples. **e**, Expression of *PAP1*, *PAP2*, *MYB113*, *MYB114* and *DFR* in 3-week-old wild-type, *s678*, *S6:S6-HA s678* and *S6:S6$^{\Delta EAR}$-HA s678* plants. Data were normalized to wild type and are means ± s.e.m.; *n* = 3 biologically independent samples. **f**, The *pap1-D* mutant rescued the anthocyanin-biosynthesis deficiency of the *max3-9* mutant. Scale bar, 1 mm. Data represent 18 independent experiments. **g**, Mutation sites in the *pap1-2* and *pap2-1* mutants. **a**–**e**, *P* values are shown; two-sided Student's *t*-test (**a**, **b**, **d**, **e**) or Tukey's HSD test (**c**).



Extended Data Fig. 8 | Feedback inhibition of *SMXL* transcription in *Arabidopsis*. **a**, SMXL6 binds to the *SMXL6* and *SMXL8* promoters in ChIP–seq (left) and ChIP–qPCR assays (right). Red lines below the peak regions show the locations of probes used in ChIP–qPCR assays and EMSAs. Chromatin from *S6:S6-HA s678* and *s678* seedlings was immunoprecipitated with anti-HA polyclonal antibodies. In ChIP–qPCR assays, the enrichment of target gene promoters is displayed as a percentage of input DNA. Values are means ± s.e.m.; $n = 4$ biologically independent samples. The *TUA2* promoter was used as a nonspecific target. **b**, Diagrams showing the constructs used in transient expression assays in protoplasts. We included the −1598 to +402 region of *SMXL6*, −1607 to +393 region of *SMXL7*, −1890 to +110 region of *SMXL8* and −2520 to +480 region of *BRC1* in *pSMXL6-LUC*, *pSMXL7-LUC*, *pSMXL8-LUC* and *pBRC1-LUC* constructs, respectively. The small black rectangle represents the EAR motif in SMXL6. Firefly LUC, firefly luciferase reporter gene; Renilla LUC, Renilla luciferase reporter gene. Firefly LUC activity was normalized against that of Renilla LUC. **c**, Transcriptional regulation by SMXL6 and SMXL6$^{\Delta EAR}$ of *SMXL6* and *SMXL8* promoters in transient expression assays. Data were normalized to samples expressing GFP and are means ± s.e.m.; $n = 3$ biologically independent samples. **d**, **e**, Expression of *SMXL6*, *SMXL7* and *SMXL8* in nonelongated axillary buds of primary rosette (RI) branches and secondary cauline (CII) branches of Col-0 and strigolactone mutants. Data were normalized to wild type and are means ± s.e.m.; $n = 3$ biologically independent samples. **a**, **c**–**e**, *P* values are shown; two-sided Student's *t*-test.



Extended Data Fig. 9 | Domain mapping of the *SMXL7* promoter bound by SMXL6 in EMSAs. a, Recombinant GST, GST–SMXL6, GST–SMXL7 and GST–SMXL8 proteins used in EMSA assays. Bands of target proteins are indicated by red arrows. Data represent four independent experiments. **b**, Direct binding of GST–SMXL6, GST–SMXL7 and GST–SMXL8 to the promoters of *SMXL6*, *7*, *8* in EMSAs. Data represent three independent experiments. **c**, **d**, Direct binding of GST–SMXL6 to the *P7-3* and *P7-3-3* regions of the *SMXL7* promoter in EMSAs. Data represent four independent experiments.



Extended Data Fig. 10 | Relationship between SMXL6, 7, 8 and SPL9, 15 in shoot branching. a, **b**, Phenotypes and quantitative analysis of wild-type (Col-0), *d14-1*, *s678*, *spl9-4 spl15-1* and *s678 spl9-4 spl15-1* plants grown for 6 weeks. Scale bar, 3 cm. Data are means ± s.e.m.; *n* = 14 plants. **c**, Phenotypes of Col-0, *s678*, *spl9-4 spl15-1* and *s678 spl9-4 spl15-1* plants grown for 3 weeks. Scale bar, 3 cm. **d**, Rosette leaf number and average primary rosette branch number per rosette leaf in Col-0 plants and the indicated mutants. Data are means ± s.e.m.; *n* = 14 plants. **e**, Expression of *BRC1* in primary rosette buds of Col-0 plants and the indicated mutants. Data were normalized to wild type and are means ± s.e.m.; *n* = 4 biologically independent samples. **b**, **d**, **e**, *P* values are shown; two-sided Student's *t*-test.

nature research

Corresponding author(s): Jiayang Li

Last updated by author(s): Mar 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	STAR (version 2.4.2a), FeatureCounts (version 1.5.0), R (version 3.3.1), and DESeq (version 1.26.0) were used in RNA-seq analysis; BWA (version 0.7.10-r789), SAMtools (version 0.1.19-44428cd), and MACS (version 1.4.2) were used in ChIP-seq analysis; EXCEL 2016 was used in two-sided Student's t-test; R (version 3.6.1) was used in Tukey's HSD test; and ImageJ (1.41o) was used in measurement of leaf length and width.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The RNA-Seq and ChIP-Seq data have been deposited in the GEO (www.ncbi.nlm.nih.gov/geo/) database under the accession numbers GSE126331 and GSE140705.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample size was chosen as large as possible and following the common practice in the field. We used 3 biological replicates for RNA-seq analysis and 2 biological replicates for ChIP-seq analysis, following the common practice in the field. In RT-qPCR and ChIP-qPCR assays, 3 or 4 biological replicates were used, following the common practice in the field.
Data exclusions	No data were excluded.
Replication	Each experiment was reproduced at least three times on separate occasions. Experimental findings were reliably reproduced.
Randomization	All samples were allocated randomly into experimental groups.
Blinding	The investigators were not blinded to allocation during experiments and outcome assessment. In order to get as objective results as possible, in multiple experiments we had other researchers repeating the experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-GFP mouse (#11814460001, clone 13.1 and 7.1, 1:2000 dilution for Western -Blotting, Roche, Lot 11751700) Anti-HA rabbit (#H6908, 1:25 dilution for ChIP, 120 μg per ChIP sample, sigma, Lot 106M4792V,) Anti-GST mouse (#G1160, clone GST-2, 1:40 dilution for EMSA assay, 3.7 μg per EMSA reaction, Sigma, Lot 035M4806V).
Validation	All antibodies used in this study were certified and validated by manufactures and vendors. The details about anti-GFP monoclonal antibody is in https://www.sigmaaldrich.com/catalog/product/roche/11814460001?lang=zh&region=CN The details about anti-HA polyclonal antibody is in https://www.sigmaaldrich.com/catalog/product/sigma/h6908?lang=zh&region=CN The details about anti-GST monoclonal antibody is in https://www.sigmaaldrich.com/catalog/product/sigma/g1160?lang=zh&region=CN

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	We have submitted raw data, the WIG and BED files to GEO. The reviewer access link is https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140705.
Files in database submission	s678 SMXL6HA_ChIP_rep1_R1.fastq.gz; s678 SMXL6HA_ChIP_rep1_R2.fastq.gz; s678_ChIP_rep1_R1.fastq.gz; s678_ChIP_rep1_R2.fastq.gz; s678 SMXL6HA_ChIP_rep2_R1.fastq.gz; s678 SMXL6HA_ChIP_rep2_R2.fastq.gz; s678_ChIP_rep2_R1.fastq.gz; s678_ChIP_rep2_R2.fastq.gz s678 SMXL6HA_Rep1.wig.gz; s678_Rep1.wig.gz; s678 SMXL6HA_Rep2.wig.gz; s678_Rep2.wig.gz; s678 SMXL6HA_Rep1_vs_s678_Rep1_peaks.bed; s678 SMXL6HA_Rep2_vs_s678_Rep2_peaks.bed

Genome browser session (e.g. UCSC)	We have uploaded the WIG and BED files to GEO.

Methodology

Replicates	We have done two biological replicates.
Sequencing depth	The sequencing depth for s678 SMXL6HA_ChIP_rep1, s678_ChIP_rep1, s678 SMXL6HA_ChIP_rep2, s678_ChIP_rep2 are 49.44x, 34.24x, 44.32 x, and 48.88 x; the total number of reads for s678 SMXL6HA_ChIP_rep1, s678_ChIP_rep1, s678 SMXL6HA_ChIP_rep2, s678_ChIP_rep2 are 82,765,244, 72,606,380, 87,725,990, and89,203,402 ; the properly mapped reads for s678 SMXL6HA_ChIP_rep1, s678_ChIP_rep1, s678 SMXL6HA_ChIP_rep2, s678_ChIP_rep2 are 72,511,020, 62,392,227,422,476, and 67,640,961; the length of reads for s678 SMXL6HA_ChIP_rep1, s678_ChIP_rep1, s678 SMXL6HA_ChIP_rep2, s678_ChIP_rep2 are 150, 150, 150, 150, and they are all paired-end.
Antibodies	Anti-HA polyclonal antibody was bought from Sigma with catalog number H6908.
Peak calling parameters	Peaks in the SMXL6:SMXL6-HA s678 transgenic line were identified by the model-based analysis software MACS (version 1.4.2) in comparison with peaks in the s678 mutant with the parameters '--mfold=10, 30', then determined by P-value less than 1e−5 and fold change more than 2.0.
Data quality	The Q30 percents for s678 SMXL6HA_ChIP_rep1_R1.fastq.gz, s678 SMXL6HA_ChIP_rep1_R2.fastq.gz, s678_ChIP_rep1_R1.fastq.gz, s678_ChIP_rep1_R2.fastq.gz, s678 SMXL6HA_ChIP_rep2_R1.fastq.gz, s678 SMXL6HA_ChIP_rep2_R2.fastq.gz, s678_ChIP_rep2_R1.fastq.gz, s678_ChIP_rep2_R2.fastq.gz are 94.82, 94.28, 91.98, 91.67, 93.14, 90.94, 93.21, and 91.22, indicating a good quality of sequencing. The numbers of peaks with P-value less than 1e-5 and enrichment above 2-fold are 1,665 and 1,401 in replicate 1 and 2 of SMXL6-HA ChIP-seq.
Software	All ChIP-seq reads were mapped to the Arabidopsis genome TAIR10 using BWA (version 0.7.10-r789) software with default parameters. Duplicated reads and reads with low mapping quality were discarded using SAMtools (version 0.1.19-44428cd) with default parameters. Peaks for SMXL6-HA were identified by the model-based analysis software MACS (version 1.4.2) in comparison with s678 with the parameters ' --mfold=10, 30'.

Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins

https://doi.org/10.1038/s41586-020-2169-0

Received: 7 February 2020

Accepted: 17 March 2020

Tommy Tsan-Yuk Lam[1,2,10], Na Jia[3,10], Ya-Wei Zhang[3,10], Marcus Ho-Hin Shum[2,10], Jia-Fu Jiang[3,10], Hua-Chen Zhu[1,2], Yi-Gang Tong[4,10], Yong-Xia Shi[5], Xue-Bing Ni[2], Yun-Shi Liao[2], Wen-Juan Li[4], Bao-Gui Jiang[3], Wei Wei[6], Ting-Ting Yuan[3], Kui Zheng[5], Xiao-Ming Cui[3], Jie Li[3], Guang-Qian Pei[3], Xin Qiang[3], William Yiu-Man Cheung[2], Lian-Feng Li[7], Fang-Fang Sun[5], Si Qin[3], Ji-Cheng Huang[5], Gabriel M. Leung[2], Edward C. Holmes[8], Yan-Ling Hu[6,9] ✉, Yi Guan[1,2] ✉ & Wu-Chun Cao[3] ✉

The ongoing outbreak of viral pneumonia in China and across the world is associated with a new coronavirus, SARS-CoV-2[1]. This outbreak has been tentatively associated with a seafood market in Wuhan, China, where the sale of wild animals may be the source of zoonotic infection[2]. Although bats are probable reservoir hosts for SARS-CoV-2, the identity of any intermediate host that may have facilitated transfer to humans is unknown. Here we report the identification of SARS-CoV-2-related coronaviruses in Malayan pangolins (*Manis javanica*) seized in anti-smuggling operations in southern China. Metagenomic sequencing identified pangolin-associated coronaviruses that belong to two sub-lineages of SARS-CoV-2-related coronaviruses, including one that exhibits strong similarity in the receptor-binding domain to SARS-CoV-2. The discovery of multiple lineages of pangolin coronavirus and their similarity to SARS-CoV-2 suggests that pangolins should be considered as possible hosts in the emergence of new coronaviruses and should be removed from wet markets to prevent zoonotic transmission.

An outbreak of serious pneumonia disease was reported in Wuhan, China, on 30 December 2019. The causative agent was soon identified as a novel coronavirus[1], which was later named SARS-CoV-2. Case numbers grew rapidly from 27 in December 2019 to 3,090,445 globally as of 30 April 2020[3], leading to the declaration of a public health emergency, and later a pandemic, by the WHO (World Health Organization). Many of the early cases were linked to the Huanan seafood market in Wuhan city, Hubei province, from where the probable zoonotic source is speculated to originate[2]. Currently, only environmental samples taken from the market have been reported to be positive for SARS-CoV-2 by the Chinese Center for Disease Control and Prevention[4]. However, as similar wet markets were implicated in the SARS outbreak of 2002–2003[5], it seems likely that wild animals were also involved in the emergence of SARS-CoV-2. Indeed, a number of mammalian species were available for purchase in the Huanan seafood market before the outbreak[4]. Unfortunately, because the market was cleared soon after the outbreak began, determining the source virus in the animal population from the market is challenging. Although a coronavirus that is closely related to SARS-CoV-2, which was sampled from a *Rhinolophus affinis* bat in Yunnan in 2013, has now been identified[6], similar viruses have not yet been detected in other wildlife species. Here we identified SARS-CoV-2-related viruses in pangolins smuggled into southern China.

We investigated the virome composition of pangolins (mammalian order Pholidota). These animals are of growing importance and interest because they are one of the most illegally trafficked mammal species: they are used as a food source and their scales are used in traditional Chinese medicine. A number of pangolin species are now regarded as critically endangered on the International Union for Conservation of Nature Red List of Threatened Species. We received frozen tissue samples (lungs, intestine and blood) collected from 18 Malayan pangolins (*Manis javanica*) during August 2017–January 2018. These pangolins were obtained during anti-smuggling operations performed by Guangxi Customs officers. Notably, high-throughput sequencing of the RNA of these samples revealed the presence of coronaviruses in 6 out of 43 samples (2 lung samples, 2 intestinal samples, 1 lung–intestine mixed sample and 1 blood sample from 5 individual pangolins; Extended Data Table 1). With the sequence read data, and by filling gaps with amplicon sequencing, we were able to obtain six complete or near complete genome sequences—denoted GX/P1E, GX/P2V, GX/P3B, GX/P4L, GX/P5E and GX/P5L—that fall into the SARS-CoV-2 lineage (within the genus *Betacoronavirus* of the *Coronaviridae*) in a

---

[1]Joint Institute of Virology (Shantou University and The University of Hong Kong), Guangdong-Hongkong Joint Laboratory of Emerging Infectious Diseases, Shantou University, Shantou, P. R. China. [2]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong, P. R. China. [3]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, P. R. China. [4]Beijing Advanced Innovation Center for Soft Matter Science and Engineering (BAIC-SM), College of Life Science and Technology, Beijing University of Chemical Technology, Beijing, P. R. China. [5]Guangzhou Customs Technology Center, Guangzhou, P. R. China. [6]Life Sciences Institute, Guangxi Medical University, Nanning, P. R. China. [7]School of Information and Management, Guangxi Medical University, Nanning, P. R. China. [8]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia. [9]Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, P. R. China. [10]These authors contributed equally: Tommy Tsan-Yuk Lam, Na Jia, Ya-Wei Zhang, Marcus Ho-Hin Shum, Jia-Fu Jiang, Yi-Gang Tong. ✉e-mail: huyanling@gxmu.edu.cn; yguan@hku.hk; caowc@bmi.ac.cn

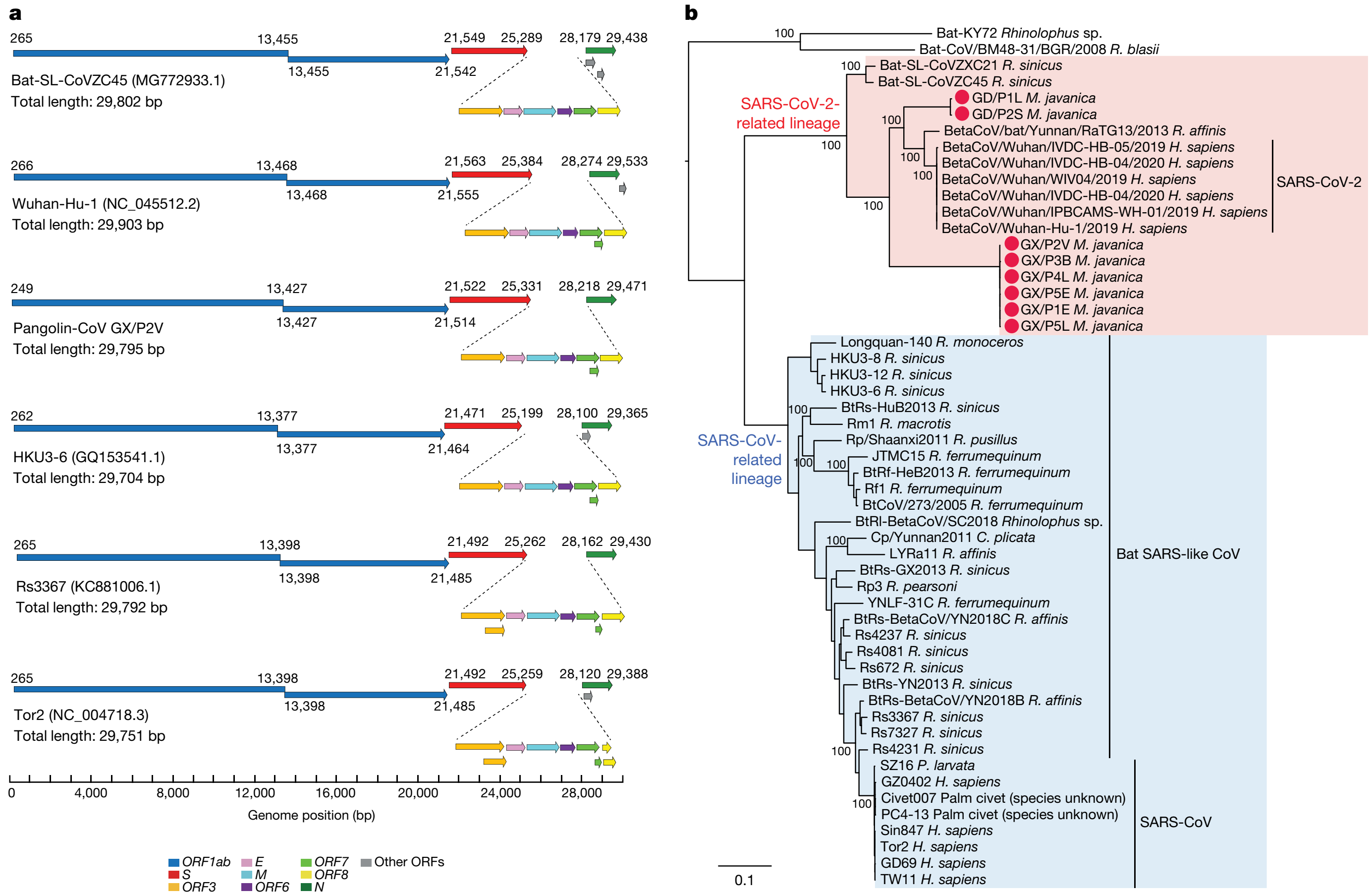


Fig. 1 | Evolutionary relationships among sequences of human SARS-CoV-2, pangolin coronaviruses and the other reference coronaviruses. a, Genome organization of coronaviruses including the pangolin coronaviruses obtained in this study, with the predicted ORFs shown in different colours (ORF1a is omitted for clarity). The pangolin coronavirus strain GX/P2V is shown with its sequence length. For comparison, the human sequences NC_045512.2 and NC_004718.3, and bat sequences MG772933.1, GQ153541.1 and KC881006.1 are included (see Extended Data Table 6 for sources). **b**, Phylogeny of the subgenus *Sarbecovirus* (genus *Betacoronavirus*; $n = 53$) estimated from the concatenated *ORF1ab*, *S*, *E*, *M* and *N* genes. Red circles indicate the pangolin coronavirus sequences generated in this study (Extended Data Table 1). GD/P1L is the consensus sequence re-assembled from previously published raw data[7]. Phylogenies were estimated using a maximum likelihood approach that used the GTRGAMMA nucleotide substitution model and 1,000 bootstrap replicates. Scientific names of the bat hosts are indicated at the end of the sequence names, and abbreviated as follows: *C. plicata*, *Chaerephon plicata*; *R. affinis*, *Rhinolophus affinis*; *R. blasii*, *Rhinolophus blasii*; *R. ferrumequinum*, *Rhinolophus ferrumequinum*; *R. monoceros*, *Rhinolophus monoceros*; *R. macrotis*, *Rhinolophus macrotis*; *R. pearsoni*, *Rhinolophus pearsoni*; *R. pusillus*, *Rhinolophus pusillus*; *R. sinicus*, *Rhinolophus sinicus*. Palm civet (*P. larvata*, *Paguma larvata*; species unspecified for Civet007 and PC4-13 sequences) and human (*H. sapiens*, *Homo sapiens*) sequences are also shown.

phylogenetic analysis (Fig. 1b). The genome sequence of the virus isolate (GX/P2V) has a very high similarity (99.83–99.92%) to the five sequences that were obtained through the metagenomic sequencing of the raw samples, and all samples have similar genomic organizations to SARS-CoV-2, with eleven predicted open-reading frames (ORFs) (Fig. 1a and Extended Data Table 2; two ORFs overlap). We were also able to successfully isolate the virus using the Vero E6 cell line (Extended Data Fig. 1). On the basis of these genome sequences, we designed primers for quantitative PCR (qPCR) detection to confirm that the raw samples were positive for coronavirus. We conducted further qPCR testing on another batch of archived pangolin samples collected between May and July 2018. Among the 19 samples (9 intestine tissues, 10 lung tissues) tested from 12 animals, 3 lung tissue samples from 3 individual pangolins were positive for coronavirus.

In addition to the animals from Guangxi, after the start of the SARS-CoV-2 outbreak researchers of the Guangzhou Customs Technology Center re-examined five archived pangolin samples (two skin swabs, two unknown tissue samples and one scale) obtained in anti-smuggling operations performed in March 2019. Following high-throughput sequencing, the scale sample was found to contain coronavirus reads, and from these data we assembled a partial genome sequence of 21,505 bp (denoted as GD/P2S), representing approximately 72% of the SARS-CoV-2 genome. Notably, this virus sequence, obtained from a pangolin scale sample, may in fact be derived from contaminants of other infected tissues. Another study of diseased pangolins in Guangdong performed in 2019 also identified viral contigs from lung samples that were similarly related to SARS-CoV-2[7]. Different assembly methods and manual curation were performed to generate a partial genome sequence that comprised 86.3% of the full-length virus genome (denoted as GD/P1L in the phylogeny shown in Fig. 1b).

These pangolin coronavirus genomes have 85.5% to 92.4% sequence similarity to SARS-CoV-2, and represent two sub-lineages of SARS-CoV-2-related viruses in the phylogenetic tree, one of which (comprising GD/P1L and GD/P2S) is very closely related to SARS-CoV-2 (Fig. 1b). It has previously been noted that members of the subgenus *Sarbecovirus* have experienced widespread recombination[8]. In support of this, a recombination analysis (Fig. 2) revealed that bat coronaviruses ZC45 and ZXC21 are probably recombinants, containing genome fragments derived from multiple SARS-CoV-related lineages (genome

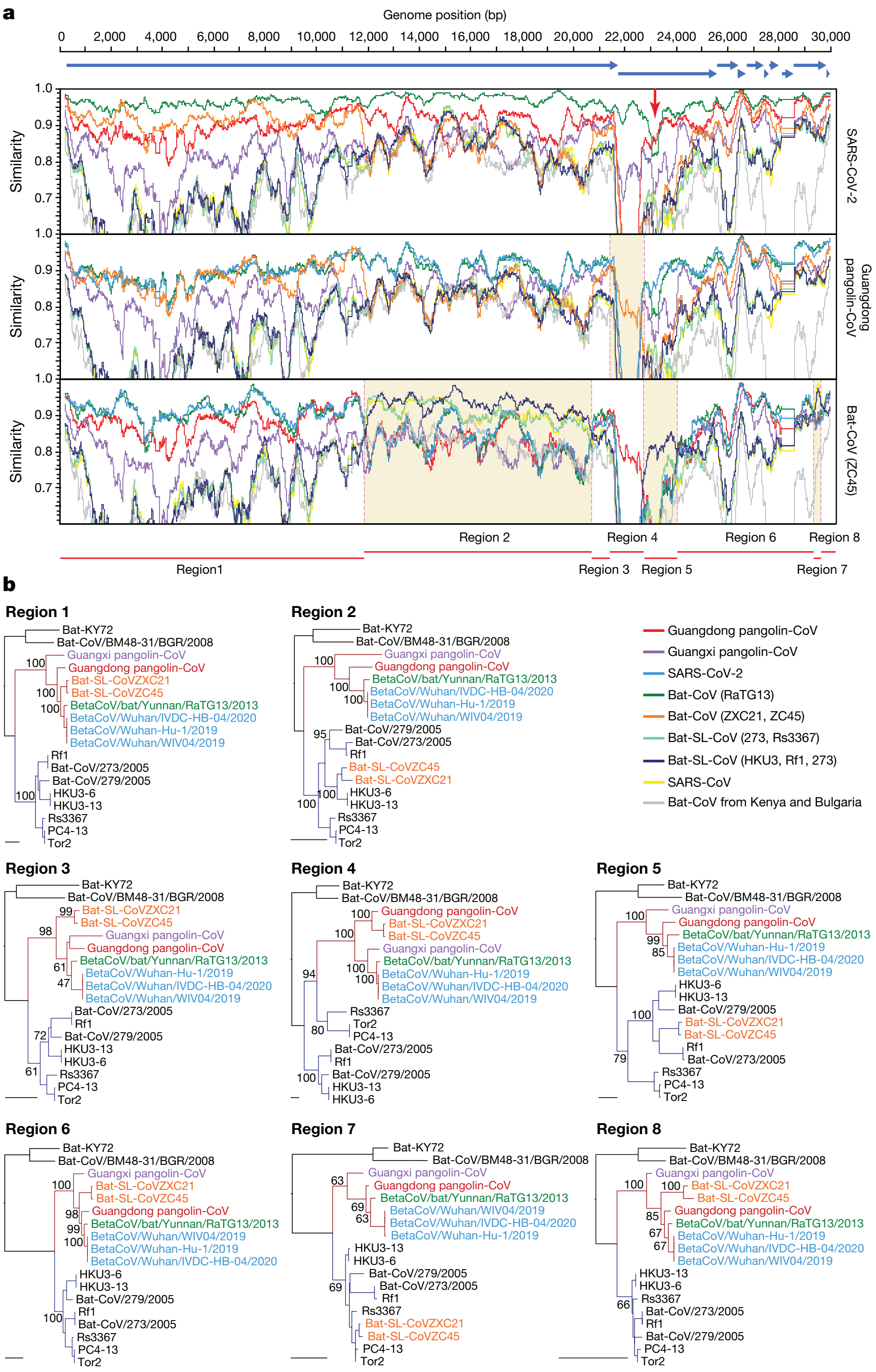
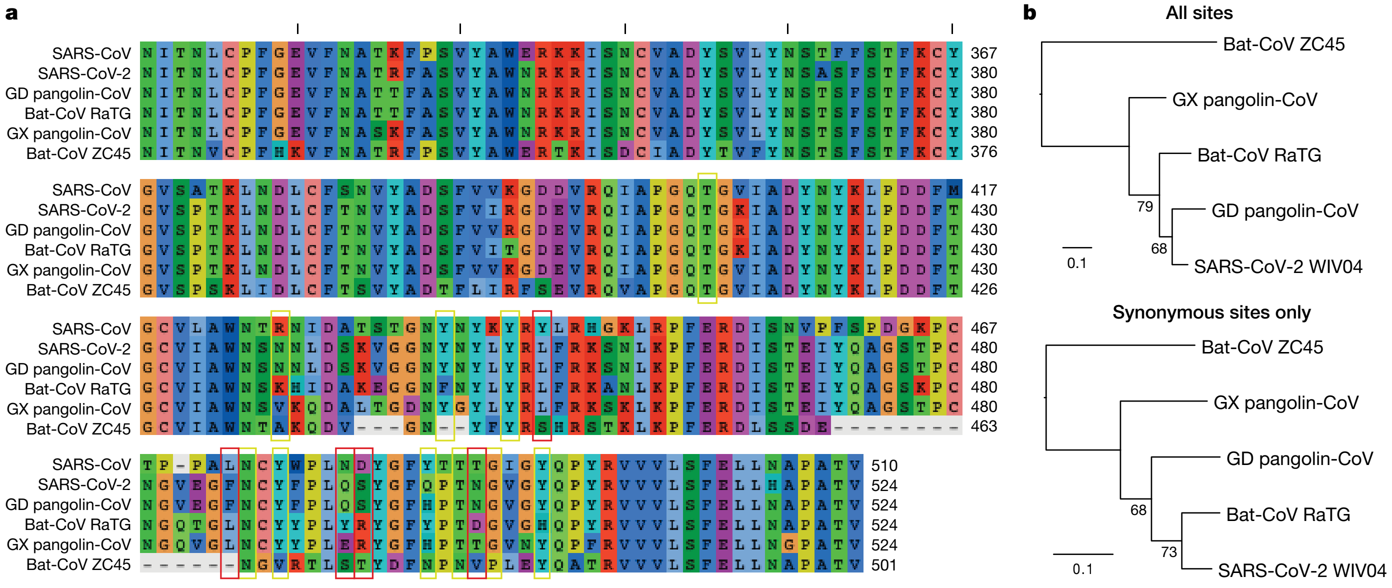


Fig. 2 | Recombination analysis. a, Sliding window analysis of changing patterns of sequence similarity between human SARS-CoV-2, pangolin and bat coronaviruses. The potential recombination breakpoints are shown in pink dash lines, and regions separated by the breakpoints are alternatively shaded in yellow. These potential breakpoints subdivide the genomes into eight regions (regions with fewer than 200 bp were omitted), indicated by the red bars at the bottom of the analysis boxes. The names of the query sequences are shown vertically to the right of the analysis boxes. The similarities to different reference sequences are indicated by different colours. Guangdong pangolin-CoV GD/P1L and pangolin-CoV GD/P2S were merged for this analysis. The blue arrows at the top indicate the position of the ORFs in the alignment. **b**, Phylogenetic trees of different genomic regions. SARS-CoV- and SARS-CoV-2-related lineages are shown in blue and red tree branches, respectively. Branch supports obtained from 1,000 bootstrap replicates are shown. Branch scale bars are shown as 0.1 substitutions per site.

regions 2, 5 and 7) as well as SARS-CoV-2-related lineages, including segments from pangolin coronaviruses (regions 1, 3, 4, 6 and 8).

More notable, however, was the observation of putative recombination signals between the pangolin coronaviruses, bat coronavirus RaTG13 and human SARS-CoV-2 (Fig. 2). In particular, SARS-CoV-2 exhibits very high sequence similarity to the Guangdong pangolin coronaviruses in the receptor-binding domain (RBD) (97.4% amino acid similarity, indicated by red arrow in Fig. 2a; the alignment is shown in Fig. 3a), even though it is most closely related to bat coronavirus RaTG13 in the remainder of the viral genome. Indeed, the Guangdong pangolin coronaviruses and SARS-CoV-2 possess identical amino acids at the five critical residues of the RBD, whereas RaTG13 only shares one amino acid with SARS-CoV-2 (residue 442, according to numbering of the human SARS-CoV[9]) and these latter two viruses have only 89.2% amino acid similarity in the RBD. Notably, a phylogenetic analysis of synonymous sites only from the RBD revealed that the topological position of the Guangdong pangolin is consistent with that of the remainder of the viral genome, rather than being the closest relative of SARS-CoV-2 (Fig. 3b). Therefore, it is possible that the amino acid similarity between the RBD of the Guangdong pangolin coronaviruses and SARS-CoV-2 is due to selectively mediated convergent evolution rather than recombination, although it is difficult to differentiate between these scenarios on the basis of the current data. This observation is consistent with the fact that the sequence similarity of ACE2 is higher between humans and pangolins (84.8%) than between humans and bats (80.8–81.4% for *Rhinolophus* sp.) (Extended Data Table 3). The occurrence of recombination and/or convergent evolution further highlights the role that intermediate animal hosts have in the emergence of viruses that can



```
       SARS-CoV NITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFKCY 367
     SARS-CoV-2 NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCY 380
GD pangolin-CoV NITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCY 380
   Bat-CoV RaTG NITNLCPFGEVFNATTFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCY 380
GX pangolin-CoV NITNLCPFGEVFNASKFASVYAWNRKRISNCVADYSVLYNSTSFSTFKCY 380
   Bat-CoV ZC45 NITNVCPFHKVFNATRFPSVYAWERTKISDCIADYTVFYNSTSFSTFKCY 376

       SARS-CoV GVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFM 417
     SARS-CoV-2 GVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFT 430
GD pangolin-CoV GVSPTKLNDLCFTNVYADSFVVRGDEVRQIAPGQTGRIADYNYKLPDDFT 430
   Bat-CoV RaTG GVSPTKLNDLCFTNVYADSFVITGDEVRQIAPGQTGKIADYNYKLPDDFT 430
GX pangolin-CoV GVSPTKLNDLCFTNVYADSFVVKGDEVRQIAPGQTGVIADYNYKLPDDFT 430
   Bat-CoV ZC45 GVSPSKLIDLCFTSVYADTFLIRFSEVRQVAPGQTGVIADYNYKLPDDFT 426

       SARS-CoV GCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPC 467
     SARS-CoV-2 GCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC 480
GD pangolin-CoV GCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC 480
   Bat-CoV RaTG GCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPC 480
GX pangolin-CoV GCVIAWNSVKQDALTGDNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPC 480
   Bat-CoV ZC45 GCVIAWNTAKQDV---GN--YFYRSHRSTKLKPFERDLSSDE-------- 463

       SARS-CoV TP-PALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATV 510
     SARS-CoV-2 NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATV 524
GD pangolin-CoV NGVEGFNCYFPLQSYGFHPTNGVGYQPYRVVVLSFELLNAPATV 524
   Bat-CoV RaTG NGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATV 524
GX pangolin-CoV NGQVGLNCYYPLERYGFHPTTGVNYQPFRVVVLSFELLNGPATV 524
   Bat-CoV ZC45 ------NGVRTLSTYDFNPNVPLEYQATRVVVLSFELLNAPATV 501
```

Fig. 3 | Analysis of the RBD sequence. a, Sequence alignment showing the RBD in human, pangolin and bat coronaviruses. The five critical residues for binding between SARS-CoV RBD and human ACE2 protein are indicated in red boxes, and ACE2-contacting residues are indicated by yellow boxes as previously described[9]. In the Guangdong pangolin-CoV sequence, the codon positions encoding the amino acids Pro337, Asn420, Pro499 and Asn519 have ambiguous nucleotide compositions, resulting in possible alternative amino acids at these sites (threonine, glycine, threonine and lysine, respectively). Sequence gaps are indicated with dashes. The short black lines at the top indicate the positions of every 10 residues. GD, Guangdong; GX, Guangxi. **b**, Phylogenetic trees of the SARS-CoV-2-related lineage estimated from the entire RBD region (top) and synonymous sites only (bottom). Branch supports obtained from 1,000 bootstrap replicates are shown. Branch scale bars are shown as 0.1 substitutions per site.

infect humans. However, all of the pangolin coronaviruses identified to date lack the insertion of a polybasic (furin-like) S1/S2 cleavage site in the spike protein that distinguishes human SARS-CoV-2 from related betacoronaviruses (including RaTG13)[10] and that may have helped to facilitate the emergence and rapid spread of SARS-CoV-2 through human populations.

To our knowledge, pangolins are the only mammals in addition to bats that have been documented to be infected by a SARS-CoV-2-related coronavirus. It is notable that two related lineages of coronaviruses are found in pangolins that were independently sampled in different Chinese provinces and that both are also related to SARS-CoV-2. This suggests that these animals may be important hosts for these viruses, which is surprising as pangolins are solitary animals that have relatively small population sizes, reflecting their endangered status[11]. Indeed, on the basis of the current data it cannot be excluded that pangolins acquired their SARS-CoV-2-related viruses independently from bats or another animal host. Therefore, their role in the emergence of human SARS-CoV-2 remains to be confirmed. In this context, it is noteworthy that both lineages of pangolin coronaviruses were obtained from trafficked Malayan pangolins, which originated from Southeast Asia, and that there is a marked lack of knowledge of the viral diversity maintained by this species in regions in which it is indigenous. Furthermore, the extent of virus transmission in pangolin populations should be investigated further. However, the repeated occurrence of infections with SARS-CoV-2-related coronaviruses in Guangxi and Guangdong provinces suggests that this animal may have an important role in the community ecology of coronaviruses.

Coronaviruses, including those related to SARS-CoV-2, are present in many wild mammals in Asia[5–7,12]. Although the epidemiology, pathogenicity, interspecies infectivity and transmissibility of coronaviruses in pangolins remains to be studied, the data presented here strongly suggests that handling these animals requires considerable caution and their sale in wet markets should be strictly prohibited. Further surveillance of pangolins in their natural environment in China and Southeast Asia are necessary to understand their role in the emergence of coronaviruses and the risk of future zoonotic transmissions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2169-0.

1. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
2. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
3. World Health Organization. *Coronavirus disease 2019 (COVID-19) Situation Report – 101* https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200430-sitrep-101-covid-19.pdf v.101 (30 April 2020).
4. Cohen, J. Mining coronavirus genomes for clues to the outbreak's origins. *Science* https://www.sciencemag.org/news/2020/01/mining-coronavirus-genomes-clues-outbreak-s-origins (31 January 2020).
5. Wang, M. et al. SARS-CoV infection in a restaurant from palm civet. *Emerg. Infect. Dis.* **11**, 1860–1865 (2005).
6. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
7. Liu, P., Chen, W. & Chen, J. P. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses* **12**, 11 (2019).
8. Hon, C. C. et al. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J. Virol.* **82**, 1819–1826 (2008).
9. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* **94**, e00127-20 (2020).
10. Coutard, B. et al. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* **176**, 104742 (2020).
11. Heinrich, S. et al. *The Global Trafficking of Pangolins: A Comprehensive Summary of Seizures and Trafficking Routes from 2010–2015* (TRAFFIC, 2017).
12. Wang, W. et al. Discovery of a highly divergent coronavirus in the Asian house shrew from China illuminates the origin of the alphacoronaviruses. *J. Virol.* **91**, e00764-17 (2017).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Ethics statement

The animals studied here were rescued and treated by the Guangxi Zhuang Autonomous Region Terrestrial Wildlife Medical-aid and Monitoring Epidemic Diseases Research Center under the ethics approval (wild animal treatment regulation No. [2011] 85). The samples were collected following the procedure guideline (Pangolins Rescue Procedure, November, 2016).

Sample collection, viral detection and sequencing of pangolins in Guangxi

We received frozen tissue samples of 18 pangolins (*M. javanica*) from Guangxi Medical University, China, that were collected between August 2017 – January 2018. These pangolins were seized by the Guangxi Customs during their routine anti-smuggling operations. All animal individuals comprised samples from multiple organs including lungs, intestine and blood, with the exception of six individuals for which only lung tissues were available, five with mixed intestine and lung tissues only, one with intestine tissues only, and one comprising two blood samples. Using the intestine–lung mixed sample we were able to isolate a novel *Betacoronavirus* using the Vero-E6 cell line (from ATCC; Extended Data Fig. 1). The cell line was subjected to species identification and authentication by microscopic morphologic evaluation and growth curve analysis, and was tested free of mycoplasma contamination. The cell line was not on the list of common misidentified cell lines by ICLAC. A High Pure Viral RNA Kit (Roche) was used for RNA extraction on all 43 samples. For RNA sequencing (GX/P2V and GX/P3B), a sequencing library was constructed using an Ion Total RNA-Seq Kit v2 (Thermo Fisher Scientific), and the library was subsequently sequenced using an Ion Torrent S5 sequencer (Thermo Fisher Scientific). For other samples, reverse transcription was performed using an SuperScript III First-Strand Synthesis System for RT–PCR (Thermo Fisher Scientific). DNA libraries were constructed using the NEBNext Ultra II DNA Library Prep Kit and sequenced on a MiSeq sequencer. The NGS (next-generation sequencing) QC Toolkit V2.3.3 was used to remove low-quality and short reads. Both BLASTn and BLASTx were used to search against a local virus database, using the data available at NCBI/GenBank. Genome sequences were assembled using the CLC Genomic Workbench v.9.0. To fill gaps in high throughput sequencing and obtain the whole viral genome sequence, amplicon primers based on the bat SARS-like coronavirus ZC45 (GenBank accession number MG772933) sequence and the coronavirus contigs obtained in the initial sequencing were designed for further amplicon-based sequencing.

A total of six samples (including the virus isolate) contained reads that matched members of the genus *Betacoronavirus* (Extended Data Table 1). We obtained near complete viral genomes from these samples (98%, compared to SARS-CoV-2), which were designated GX/P1E, GX/P2V, GX/P3B, GX/P4L, GX/P5E and GX/P5L. Their average sequencing coverage ranged from approximately 8.4X to 8,478X (Extended Data Fig. 2a-f). On the basis of these genome sequences, we designed primers for qPCR to confirm the positivity of the original tissue samples (Extended Data Table 4). This revealed an original lung tissue sample that was also qPCR positive, in addition to the six original samples with coronavirus reads. We further tested an additional 19 samples (nine intestine tissues and ten lung tissues), from 12 smuggled pangolins sampled between May–July 2018 by the group from Guangxi Medical University. The genome sequences of GX/P1E, GX/P2V, GX/P3B, GX/P4L, GX/P5E and GX/P5L have been submitted to GISAID database and assigned accession numbers EPI_ISL_410538 - EPI_ISL_410543.

Sample collection, viral detection and sequencing of pangolins in Guangdong

After the start of the SARS-CoV-2 outbreak, the Guangzhou Customs Technology Center re-examined their five archived pangolin samples (two skin swabs, two unknown tissue and one scale) obtained in anti-smuggling operations undertaken in March 2019. RNA was extracted from all five samples (Qiagen), and was subjected to high-throughput RNA sequencing on the Illumina HiSeq platform by Vision medicals. The scale sample was found to contain coronavirus reads using a BLAST-based approach. These reads were quality assessed, cleaned and assembled into contigs by both de novo (MEGAHIT v1.1.3[13]) and using reference (BWA v0.7.13[14]) assembly methods, using BetaCoV/Wuhan/WIV04/2019 as a reference. The contigs were combined, and approximately 72% of the coronavirus genome (21,505 bp) was obtained. This sequence has about 6.6× sequencing coverage (Extended Data Fig. 2g) and denoted pangolin-CoV GD/P2S. This sequence has been deposited on GISAID with accession number EPI_ISL_410544.

A recently published meta-transcriptomic study of pangolins[7] deposited 21 RNA-seq raw files on the SRA database (https://www.ncbi.nlm.nih.gov/sra). We screened these raw read files using BLAST methods and found that five (SRR10168374, SRR10168376, SRR10168377, SRR10168378 and SRR10168392) contained reads that mapped to SARS-CoV-2. These reads were subjected to quality assessment, cleaning and then de novo assembly using MEGAHIT[13] and reference assembly using BWA[14]. These reads were then merged and curated in a pileup alignment file to obtain the consensus sequences. This combined consensus sequence is 25,753 bp in length (about 86.3% of BetaCoV/Wuhan/WIV04/2019; about 6.9× coverage) and denoted pangolin-CoV GD/P1L (available in the Supplementary Information Dataset). Notably, it has 66.8% overlap and a sequence identity of 99.79% with the GD/P2S sequence. As the genetic distance between these viruses is very low, for the recombination analysis we merged the GD/P1L and GD/P2S sequences into a single consensus sequence to minimize gap regions within any sequences.

The viral genome organizations of the Guangxi and Guangdong pangolin coronaviruses were similar to SARS-CoV-2. They possessed nine non-overlapping open reading frames (ORFs) plus two overlapping ORFs, and shared the same gene order of ORF1ab replicase, envelope glycoprotein spike (S), envelope (E), membrane (M), nucleocapsid (N), plus other predicted ORFs. A detailed comparison of the ORF length and similarity with SARS-CoV-2 and bat coronavirus RaTG13 is provided in Extended Data Table 2.

Sequence, phylogenetic and recombination analyses

The human SARS-CoV-2 and bat RaTG13 coronavirus genome sequences were downloaded from Virological.org (http://virological.org) and the GISAID (https://www.gisaid.org) databases in January 2020, with the data kindly shared by the submitters (Extended Data Table 5). Other coronaviruses (subgenus *Sarbecovirus*) were downloaded from GenBank (Extended Data Table 6) and compared to those obtained here. We constructed a multiple sequence alignment of their complete genomes and individual genes using MAFFT v.7.273[15]. Maximum likelihood phylogenies were estimated using RAxML v.8.2.12[16] from 100 inferences, using the GTRGAMMA model of nucleotide substitution with 1,000 bootstrap replicates. To investigate potential recombination events, we used SimPlot v.3.5.1[17] to conduct a window sliding analysis to determine the changing patterns of sequence similarity and phylogenetic clustering between the query and the reference sequences. A full plot for the recombination analysis is provided in Extended Data Fig. 3. We also examined phylogenetic clusters performed directly from the multiple sequence alignment. Maximum likelihood trees were estimated from each window extraction (that is, genome regions 1 to 8) using RAxML as described above.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data that support the findings of this study have been deposited in the GISAID database (https://www.gisaid.org) with accession numbers EPI_ISL_410538–EPI_ISL_410544 and the SRA database under BioProject accession number PRJNA606875. The data are also available as Supplementary Information.

13. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
14. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
15. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res*. **47**, W5–W10 (2019).
16. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
17. Lole, K. S. et al. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol*. **73**, 152–160 (1999).
18. He, R. et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun*. **316**, 476–483 (2004).
19. Snijder, E. J. et al. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol*. **331**, 991–1004 (2003).
20. Marra, M. A. et al. The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404 (2003).
21. Song, H. D. et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl Acad. Sci. USA* **102**, 2430–2435 (2005).
22. Han, Y. et al. Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Front. Microbiol*. **10**, 1900 (2019).
23. Tao, Y. & Tong, S. Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. *Microbiol. Resour. Announc*. **8**, e00548-19 (2019).
24. Hu, D. et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect*. **7**, 1–10 (2018).
25. Hu, B. et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog*. **13**, e1006698 (2017).
26. Wu, Z. et al. ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis*. **213**, 579–583 (2016).
27. Wu, Z. et al. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *ISME J*. **10**, 609–620 (2016).
28. Yang, L. et al. Novel SARS-like betacoronaviruses in bats, China, 2011. *Emerg. Infect. Dis*. **19**, 989–991 (2013).
29. Xu, L. et al. Detection and characterization of diverse alpha- and betacoronaviruses from bats in China. *Virol. Sin*. **31**, 69–77 (2016).
30. He, B. et al. Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China. *J. Virol*. **88**, 7070–7082 (2014).
31. Ge, X. Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
32. Li, W. et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
33. Drexler, J. F. et al. Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *J. Virol*. **84**, 11336–11349 (2010).
34. Lau, S. K. et al. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related *Rhinolophus* bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. *J. Virol*. **84**, 2808–2819 (2010).
35. Yuan, J. et al. Intraspecies diversity of SARS-like coronaviruses in *Rhinolophus sinicus* and its implications for the origin of SARS coronaviruses in humans. *J. Gen. Virol*. **91**, 1058–1062 (2010).
36. Guan, Y. et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
37. Tang, X. C. et al. Prevalence and genetic diversity of coronaviruses in bats from China. *J. Virol*. **80**, 7481–7490 (2006).
38. Yeh, S. H. et al. Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. *Proc. Natl Acad. Sci. USA* **101**, 2542–2547 (2004).
39. Vega, V. B. et al. Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infect. Dis*. **4**, 32 (2004).

Acknowledgements We thank the staff of the Guangxi and Guangdong Custom Bureau for their laborious anti-smuggling operations and all of the scientists who shared their genomic sequences of the coronaviruses used in this study. The computations were performed using research computing facilities offered by Information Technology Services, the University of Hong Kong. This work was supported by research grants from The Natural Science Foundation of China (NSFC; 81621005), the State Key Research Development Program of China (2019YFC1200401), NSFC Excellent Young Scientists Fund (Hong Kong and Macau) (31922087), National Key Plan for Scientific Research and Development of China (2016YFD0500302 and 2017YFE0190800), funding for Guangdong–Hongkong–Macau Joint Laboratory (2019B121205009), Li Ka Shing Foundation, National Institutes of Health (HHSN272201400006C), Guangxi Scientific and Technological Research (2020AB39264) and Guangxi Medical University Training Program for Distinguished Young Scholars, and the Australian Research Council (FL170100022).

Author contributions W.-C.C., N.J., J.-C.H., Y.G. and Y.-L.H. designed and supervised the research. J.-F.J., B.-G.J., W.W., T.-T.Y., K.Z., L.-F.L. and X.-M.C. collected samples. Y.-W.Z., Y.-X.S., W.-J.L., W.W., T.-T.Y., J.L. and L.-F.L. prepared materials for sequencing. Y.-W.Z., Y.-X.S., G.-Q.P., X.Q., F.-F.S. and S.Q. performed genome sequencing. Y.-W.Z., M.H.-H.S., X.-B.N. and T.T.-Y.L. performed genome assembly and annotation. Y.-G.T., T.T.-Y.L., M.H.-H.S., Y.-W.Z., X.-B.N., E.C.H., Y.-S.L. and N.J. performed the genome analysis and interpretation. T.T.-Y.L., N.J., E.C.H. and W.-C.C. wrote the paper. All authors took part in data interpretation and edited the paper.

Competing interests The authors declare no competing interests.

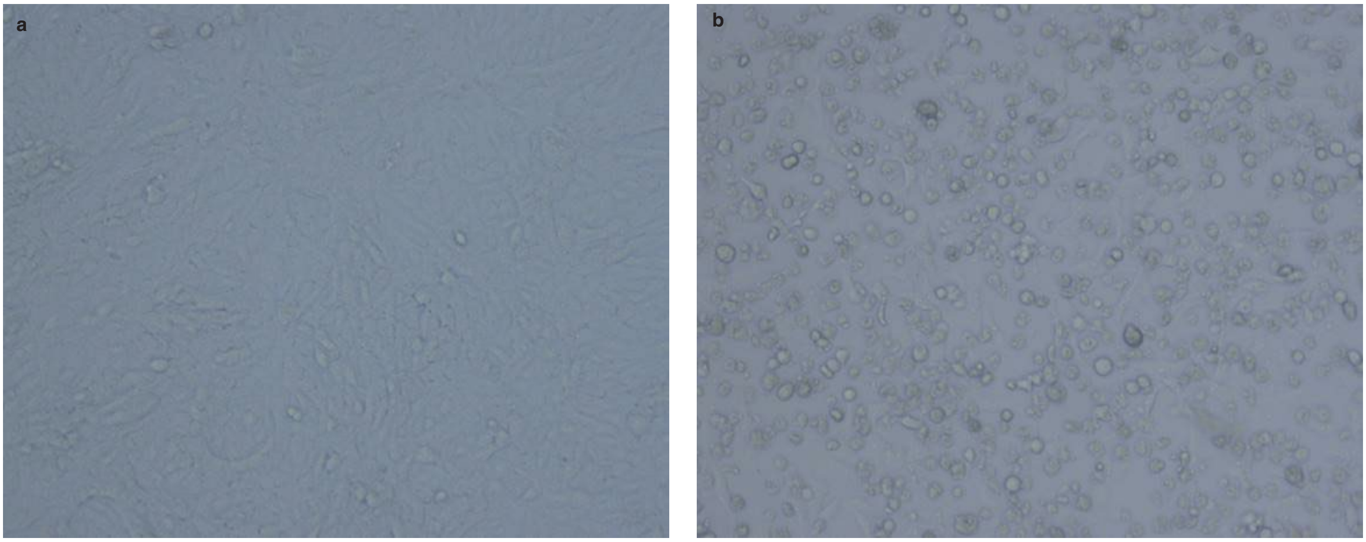
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2169-0.

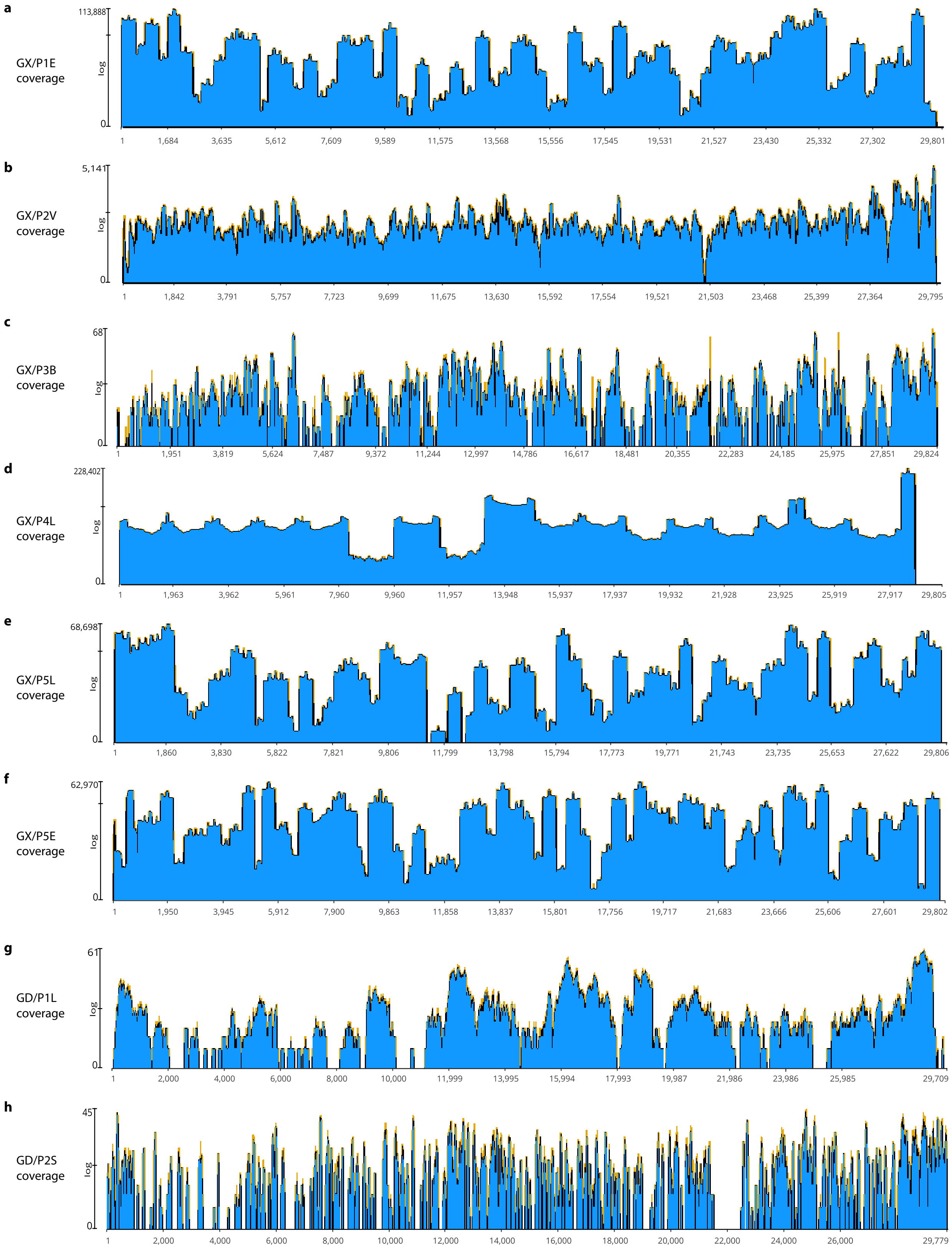
Correspondence and requests for materials should be addressed to Y.-L.H., Y.G. or W.-C.C.

Peer review information *Nature* thanks Paul Kellam, W. Ian Lipkin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

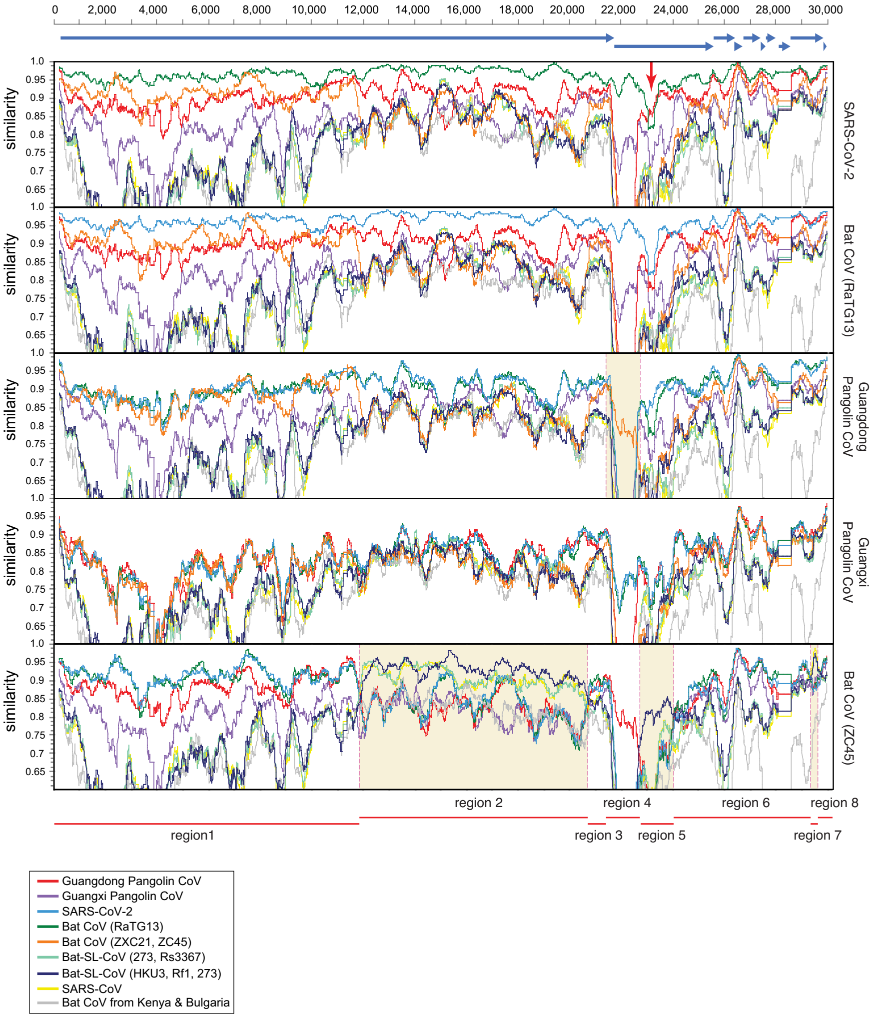
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Microscopy image of the cytopathic effect of the virus in Vero E6 cells. a, Negative control. Uninfected cells of the Vero E6 cell line. **b**, Cytopathic effect seen in viral culture (five days after inoculation). The experiment was performed twice independently in two laboratories and produced similar results.



Extended Data Fig. 2 | Read coverage depth of each pangolin coronavirus analysed in this study. **a**, GX/P1E. **b**, GX/P2V. **c**, GX/P3B. **d**, GX/P4L. **e**, GX/P5L. **f**, GX/P5E. **g**, GD/P1L. **h**, GD/P2S.



Extended Data Fig. 3 | Recombination analysis of all members of the SARS-CoV-2-related lineages. Sliding window analysis of changes in the patterns of sequence similarity between human SARS-CoV-2, and pangolin and bat coronaviruses as described further in Fig. 2a. Sequence similarity patterns of Bat-CoV (RaTG13) and Guangxi pangolin-CoV are shown in this figure but not in Fig. 2a.

Extended Data Table 1 | High-throughput sequencing results of the pangolin samples with coronavirus reads

Source location	Sample type	Sample number	Accession IDs of consensus sequence / read data
Guangxi	Intestine	GX/P1E	EPI_ISL_410539 / SAMN14115945
Guangxi	Virus isolate from intestine-lung mixed samples	GX/P2V	EPI_ISL_410542 / SAMN14115940
Guangxi	Blood	GX/P3B	EPI_ISL_410543 / SAMN14115941
Guangxi	Lung	GX/P4L	EPI_ISL_410538 / SAMN14115942
Guangxi	Intestine	GX/P5E	EPI_ISL_410541 / SAMN14115943
Guangxi	Lung	GX/P5L	EPI_ISL_410540 / SAMN14115944
Guangdong	Scale	GD/P2S	EPI_ISL_410544 / SAMN14116618

Sequencing reads have been deposited in the SRA database under BioProject accession number PRJNA606875.

Extended Data Table 2 | Genomic comparison of SARS-CoV-2 with bat-CoV RaTG13, Guangdong pangolin-CoV and Guangxi pangolin-CoV

	Bat-Cov RaTG13#			Guangdong pangolin CoV #			Guangxi pangolin CoV #		
	Length bat/ SARS-CoV-2 (bp)	nt Identity %	aa Identity %	Length GD/ SARS-CoV-2 (bp)	nt Identity %	aa Identity %	Length GX/ SARS-CoV-2 (bp)	nt Identity %	aa Identity %
ORF1ab	21287/21290	96.5	98.6	20076*/21290	90.7	97.1	21266/21290	84.9	92.5
S	3810/3822	93.1	97.7	3548*/3822	84.9	90.7	3804/3822	83.6	92.6
ORF3a	828/828	96.3	97.8	828/828	93.6	97.4	828/828	87.0	89.3
E	228/228	99.6	100	228/228	99.1	100	228/228	97.4	100
M	666/669	95.9	100	669/669	93.4	98.6	669/669	91.3	98.2
ORF6	186/186	98.4	100	186/186	95.7	96.6	186/186	90.9	95.0
ORF7a	366/366	95.6	97.5	366/366	93.4	97.5	366/366	86.6	87.7
ORF8	366/366	97.0	94.9	366/366	92.3	94.9	366/366	80.6	86.8
N	1260/1260	96.9	99.0	1260/1260	96.2	97.8	1254/1260	91.4	94.3

*Partial sequence.

#Wuhan-Hu-1 SARS-CoV-2 (NC_045512.2) was used for comparison with bat-CoV RaTG13 (EPI_ISL_402131), Guangdong pangolin-CoV (merge of GD/P1L and GD/P2S) and Guangxi pangolin-CoV (GX/P5L).

Extended Data Table 3 | Sequence similarity of amino acid sequences of ACE2 between humans, pangolins and bats

	Homo sapiens	*Manis javanica*	*Rhinolophus sinicus*	*Rhinolophus pearsonii*	*Rhinolophus ferrumequinum*
Homo sapiens	100%				
Manis javanica	84.85%	100%			
Rhinolophus sinicus	80.75%	82.86%	100%		
Rhinolophus pearsonii	81.37%	82.98%	94.41%	100%	
Rhinolophus ferrumequinum	81.24%	82.98%	93.04%	92.42%	100%
Rhinolophus macrotis	80.87%	83.73%	95.78%	94.91%	92.55%

Extended Data Table 4 | Primers used for qPCR detection of pangolin-associated coronaviruses

pCov-Forward	AGGTGACGAGGTTAGACAAATAG
pCov-Reverse	CCAAGCAATAACACAACCAGTAA
pCov-Probe	ACCCGGACAAACTGGTGTTATTGCT

Extended Data Table 5 | Previously obtained SARS-CoV-2 genome sequences

Accession ID	Virus name	Location	Collection date	Originating lab	Submitting lab	Authors
Virologal.org sequence (NC_045512.2)	BetaCoV/Wuhan-Hu-1/2019	China / Wuhan	2019-12	National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)	National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)	Zhang,Y.-Z., Wu,F., Chen,Y.-M., Pei,Y.-Y., Xu,L., Wang,W., Zhao,S., Yu,B., Hu,Y., Tao,Z.-W., Song,Z.-G., Tian,J.-H., Zhang,Y.-L., Liu,Y., Zheng,J.-J., Dai,F.-H., Wang,Q.-M., She,J.-L. and Zhu,T.-Y.
EPI_ISL_402131	BetaCoV/bat/Yunnan/RaTG13/2013	China / Yunnan Province / Pu'er City	2013-07-24	Wuhan Institute of Virology, Chinese Academy of Sciences	Wuhan Institute of Virology, Chinese Academy of Sciences	Yan Zhu, Ping Yu, Bei Li, Ben Hu, Hao-Rui Si, Xing-Lou Yang, Peng Zhou, Zheng-Li Shi
EPI_ISL_402121	BetaCoV/Wuhan/IVDC-HB-05/2019	China / Hubei Province / Wuhan City	2019-12-30	National Institute for Viral Disease Control and Prevention, China CDC	National Institute for Viral Disease Control and Prevention, China CDC	Wenjie Tan，Xuejun Ma，Xiang Zhao，Wenling Wang，Yongzhong Jiang，Roujian Lu，Ji Wang，Peihua Niu, Weimin Zhou, Faxian Zhan，Weifeng Shi，Baoying Huang，Jun Liu，Li Zhao，Yao Meng，Fei Ye，Na Zhu, Xiaozhou He，Peipei Liu, Yang Li，Jing Chen，Wenbo Xu，George F. Gao，Guizhen Wu
EPI_ISL_402120	BetaCoV/Wuhan/IVDC-HB-04/2020	China / Hubei Province / Wuhan City	2020-01-01	National Institute for Viral Disease Control and Prevention, China CDC	National Institute for Viral Disease Control and Prevention, China CDC	Wenjie Tan，Xiang Zhao，Wenling Wang，Xuejun Ma，Yongzhong Jiang，Roujian Lu，Ji Wang，Weimin Zhou，Peihua Niu，Peipei Liu，Faxian Zhan，Weifeng Shi，Baoying Huang，Jun Liu，Li Zhao，Yao Meng，Xiaozhou He，Fei Ye，Na Zhu，Yang Li，Jing Chen，Wenbo Xu，George F. Gao，Guizhen Wu
EPI_ISL_402124	BetaCoV/Wuhan/WIV04/2019	China / Hubei Province / Wuhan City	2019-12-30	Wuhan Jinyintan Hospital	Wuhan Institute of Virology, Chinese Academy of Sciences	Peng Zhou, Xing-Lou Yang, Ding-Yu Zhang, Lei Zhang, Yan Zhu, Hao-Rui Si, Zhengli Shi
EPI_ISL_402123	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	China / Hubei Province / Wuhan City	2019-12-24	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Lili Ren, Jianwei Wang, Qi Jin, Zichun Xiang, Zhiqiang Wu, Chao Wu, Yiwei Liu

SARS-CoV-2 genome sequences are available at virological.org and in the GISAID (https://www.gisaid.org) databases.

Extended Data Table 6 | GenBank accession numbers of coronavirus sequences used in this study

Accession ID	Strain name	Host	Publication
NC_004718.3	Tor2	*Homo sapiens*	He *et al.* Biochem Biophys Res Commun. 316(2):476-83 (2004) [18] Snijder *et al.* J. Mol. Biol. 331 (5), 991-1004 (2003) [19] Marra *et al.* Science 300 (5624), 1399-1404 (2003) [20]
AY313906.1	GD69	*Homo sapiens*	Song *et al.* Proc. Natl. Acad. Sci. U. S. A. 102(7):2430-5 (2005) [21]
MK211377.1	BtRs-BetaCoV/ YN2018C	*Rhinolophus affinis*	Han *et al.* Front Microbiol. 10:1900 (2019)[22]
MK211376.1	BtRs-BetaCoV/ YN2018B	*Rhinolophus affinis*	Han *et al.* Front Microbiol. 10:1900 (2019) [22]
MK211374.1	BtRl-BetaCoV/ SC2018	*Rhinolophus sp.*	Han *et al.* Front Microbiol. 10:1900 (2019) [22]
KY352407.1	BtKY72	*Rhinolophus sp.*	Tao *et al.* Microbiol Resour Announc 8 (28), e00548-19 (2019) [23]
MG772934.1	bat-SL-CoVZXC21	*Rhinolophus sinicus*	Hu et al. Emerg Microbes Infect. 12;7(1):154 (2018) [24]
MG772933.1	bat-SL-CoVZC45	*Rhinolophus sinicus*	Hu et al. Emerg Microbes Infect. 12;7(1):154 (2018) [24]
KY417151.1	Rs7327	*Rhinolophus sinicus*	Hu *et al.* PLoS Pathog. 13 (11), e1006698 (2017) [25]
KY417147.1	Rs4237	*Rhinolophus sinicus*	Hu *et al.* PLoS Pathog. 13 (11), e1006698 (2017) [25]
KY417146.1	Rs4231	*Rhinolophus sinicus*	Hu *et al.* PLoS Pathog. 13 (11), e1006698 (2017) [25]
KY417143.1	Rs4081	*Rhinolophus sinicus*	Hu *et al.* PLoS Pathog. 13 (11), e1006698 (2017) [25]
KJ473816.1	BtRs-YN2013	*Rhinolophus sinicus*	Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
KJ473815.1	BtRs-GX2013	*Rhinolophus sinicus*	Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
KJ473814.1	BtRs-HuB2013	*Rhinolophus sinicus*	Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
KJ473812.1	BtRf-HeB2013	*Rhinolophus ferrumequinum*	Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
JX993988.1	Cp/Yunnan2011	*Chaerephon plicata*	Yang *et al.* Emerging Infect. Dis. 19 (6) (2013) [28] Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
JX993987.1	Rp/Shaanxi2011	*Rhinolophus pusillus*	Yang *et al.* Emerging Infect. Dis. 19 (6) (2013) [28] Wu *et al.* J. Infect. Dis. 213 (4), 579-583 (2016) [26] Wu *et al.* ISME J 10 (3), 609-620 (2016) [27]
KU182964.1	JTMC15	*Rhinolophus ferrumequinum*	Xu *et al.* Virol Sin 31 (1), 69-77 (2016) [29]
KP886808.1	YNLF_31C	*Rhinolophus ferrumequinum*	*Journal information is not available in the GenBank record*
KF569996.1	LYRa11	*Rhinolophus affinis*	He *et al.* J. Virol. 88 (12), 7070-7082 (2014) [30]
KC881006.1	Rs3367	*Rhinolophus sinicus*	Ge *et al.* Nature 503, 535-538 (2013) [31]
DQ412043.1	Rm1	*Rhinolophus macrotis*	Li *et al.* Science 310 (5748), 676-679 (2005) [32]
DQ412042.1	Rf1	*Rhinolophus ferrumequinum*	Li *et al.* Science 310 (5748), 676-679 (2005) [32]
GU190215.1	BtCoV/BM48-31/BGR/2008	*Rhinolophus blasii*	Drexler *et al.* J. Virol. 84 (21), 11336-11349 (2010) [33]
GQ153547.1	HKU3-12	*Rhinolophus sinicus*	Lau *et al.* J. Virol. 84 (6), 2808-2819 (2010) [34]
GQ153543.1	HKU3-8	*Rhinolophus sinicus*	Lau *et al.* J. Virol. 84 (6), 2808-2819 (2010) [34]
GQ153541.1	HKU3-6	*Rhinolophus sinicus*	Lau *et al.* J. Virol. 84 (6), 2808-2819 (2010) [34]
FJ588686.1	Rs672	*Rhinolophus sinicus*	Yuan *et al.* J. Gen. Virol. 91 (PT 4), 1058-1062 (2010) [35]
DQ071615.1	Rp3	*Rhinolophus pearsoni*	Li *et al.* Science 310 (5748), 676-679 (2005) [32]
AY304488.1	SZ16	*Paguma larvata*	Guan *et al.* Science 302 (5643), 276-278 (2003) [36]
DQ648856.1	BtCoV/273/2005	*Rhinolophus ferrumequinum*	Tang *et al.* J. Virol. 80 (15), *7481-7490* (2006) [37]
AY572034.1	civet007	Palm civet (species unspecified)	Wang *et al.* Emerging Infect. Dis. 11 (12), 1860-1865 (2005) [5]
AY502924.1	TW11	*Homo sapiens*	Yeh *et al.* Proc. Natl. Acad. Sci. U.S.A. 101 (8), 2542-2547 (2004) [38]
AY613948.1	PC4_13	Palm civet (species unspecified)	Song *et al.* Proc. Natl. Acad. Sci. U.S.A. 102 (7), 2430-2435 (2005) [21]
AY613947.1	GZ0402	*Homo sapiens*	Song *et al.* Proc. Natl. Acad. Sci. U.S.A. 102 (7), 2430-2435 (2005) [21]
AY559095.1	Sin847	*Homo sapiens*	Vega *et al.* BMC Infect. Dis. 4, 32 (2004) [39]
KF294457.1	Longquan-140	*Rhinolophus monoceros*	*Journal information is not available in the GenBank record*
DQ648857.1	BtCoV/279/2005	*Rhinolophus macrotis*	Tang *et al.* J. Virol. 80 (15), 7481-7490 (2006) [37]

Sequences were published previously[18–39].

nature research

Corresponding author(s): Yi Guan
Last updated by author(s): Mar 3, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Reference genome sequence data were downloaded from GenBank, Virological.org and GISAID using the web interface.
Data analysis	Software used: CLC Genomic Workbench v9.0, BLAST v2.3.0+, BWA v0.7.13, MEGAHIT v1.1.3, MAFFT v7.273, PhyML v3.1, Simplot v3.5.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data that support the findings of this study have been deposited in GISAID database with accession numbers EPI_ISL_410538 - EPI_ISL_410544 and to the SRA database under BioProject number PRJNA606875, and are available in the supplementary information file of this paper.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We screened all relevant pangolin samples that are available to us in the study period. Among the 43 Guangxi pangolin samples (18 pangolin individuals), 6 samples (5 pangolin individuals) were found with SARS-CoV-2 related coronavirus by sequencing. Among the 5 Guangdong pangolin samples, 1 was found with SARS-CoV-2 related coronavirus by sequencing. All these coronaviruses shared >99.7% genomic similarity to either some of them among themselves or the coronavirus found in previous study. Therefore, such sample size is sufficient for the discovery of SARS-CoV-2 related coronavirus in the pangolins in our conditions.
Data exclusions	No data were excluded.
Replication	qPCR was also applied on the same sets of samples that have been examined by metatranscriptomic sequencing, as to verify the presence of pangolin coronavirus sequence indicated by sequencing.
Randomization	There was no separation of experimental groups in the study, hence no randomization.
Blinding	There was no separation of experimental groups in the study, hence no blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Vero E6 cells from ATCC.
Authentication	All Vero E6 cells were from ATCC with authentication. The authentication was performed by morphology check under microscopes and growth curve analysis.
Mycoplasma contamination	We confirm that all cells were tested as mycoplasma negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	No laboratory animals were involved in the study.
Wild animals	Thirty-five (18+12+5) pangolins were seized during routine anti-smuggling operations, and unfortunately dead for unknown reason in the rescue centre. Samples were then collected from them.
Field-collected samples	No field-collected samples were involved in the study.
Ethics oversight	The animals were rescued and treated by the Guangxi Zhuang Autonomous Region Terrestrial Wildlife Medical-aid and Monitoring Epidemic Diseases Research Center under the ethics approval (wild animal treatment regulation No. [2011] 85). The samples were collected following the procedure guideline (Pangolins Rescue Procedure, November, 2016).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins

https://doi.org/10.1038/s41586-020-2313-x

Received: 16 February 2020

Accepted: 28 April 2020

Kangpeng Xiao[1,2,7], Junqiong Zhai[3,7], Yaoyu Feng[1,2], Niu Zhou[3], Xu Zhang[1,2], Jie-Jian Zou[4], Na Li[1,2], Yaqiong Guo[1,2], Xiaobing Li[1], Xuejuan Shen[1], Zhipeng Zhang[1], Fanfan Shu[1,2], Wanyi Huang[1,2], Yu Li[5], Ziding Zhang[5], Rui-Ai Chen[1,6], Ya-Jiang Wu[3], Shi-Ming Peng[3], Mian Huang[3], Wei-Jun Xie[3], Qin-Hui Cai[3], Fang-Hui Hou[4], Wu Chen[3]✉, Lihua Xiao[1,2]✉ & Yongyi Shen[1,2]✉

The current outbreak of coronavirus disease-2019 (COVID-19) poses unprecedented challenges to global health[1]. The new coronavirus responsible for this outbreak—severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—shares high sequence identity to SARS-CoV and a bat coronavirus, RaTG13[2]. Although bats may be the reservoir host for a variety of coronaviruses[3,4], it remains unknown whether SARS-CoV-2 has additional host species. Here we show that a coronavirus, which we name pangolin-CoV, isolated from a Malayan pangolin has 100%, 98.6%, 97.8% and 90.7% amino acid identity with SARS-CoV-2 in the E, M, N and S proteins, respectively. In particular, the receptor-binding domain of the S protein of pangolin-CoV is almost identical to that of SARS-CoV-2, with one difference in a noncritical amino acid. Our comparative genomic analysis suggests that SARS-CoV-2 may have originated in the recombination of a virus similar to pangolin-CoV with one similar to RaTG13. Pangolin-CoV was detected in 17 out of the 25 Malayan pangolins that we analysed. Infected pangolins showed clinical signs and histological changes, and circulating antibodies against pangolin-CoV reacted with the S protein of SARS-CoV-2. The isolation of a coronavirus from pangolins that is closely related to SARS-CoV-2 suggests that these animals have the potential to act as an intermediate host of SARS-CoV-2. This newly identified coronavirus from pangolins—the most-trafficked mammal in the illegal wildlife trade—could represent a future threat to public health if wildlife trade is not effectively controlled.

As coronaviruses are common in mammals and birds[5], we used the whole-genome sequence of SARS-CoV-2 (strain WHCV; GenBank accession number MN908947) in a Blast search of SARS-related coronavirus sequences in available mammalian and avian viromic, metagenomic and transcriptomic data. We identified 34 closely related contigs in a set of viral metagenomes from pangolins (Extended Data Table 1), and therefore focused our subsequent search on SARS-related coronaviruses in pangolins.

We obtained the lung tissues from 4 Chinese pangolins (*Manis pentadactyla*) and 25 Malayan pangolins (*Manis javanica*) from a wildlife rescue centre during March–August 2019, and analysed them for SARS-related coronaviruses using reverse-transcription polymerase chain reaction (RT–PCR) with primers that target a conservative region of betacoronaviruses. RNA from 17 of the 25 Malayan pangolins generated the expected PCR product, whereas RNA from the Chinese pangolins did not amplify. The virus-positive Malayan pangolins were all from the first transport. These pangolins were brought into the rescue centre at the end of March, and gradually showed signs of respiratory disease, including shortness of breath, emaciation, lack of appetite, inactivity and crying. Furthermore, 14 of the 17 pangolins that tested positive for viral RNA died within one and half months of testing. Plasma samples of four PCR-positive and four PCR-negative Malayan pangolins were used in the detection of IgG and IgM antibodies against SARS-CoV-2 using a double-antigen sandwich enzyme-linked immunosorbent assay (ELISA). One of the PCR-positive sample reacted strongly, showing an optical density at 450 nm (OD_{450}) value of 2.17 (cut-off value = 0.11) (Extended Data Table 2). The plasma remained positive at the dilution of 1:80, which suggests that the pangolin was naturally infected with a virus similar to SARS-CoV-2. The other three PCR-positive pangolins had no detectable antibodies against SARS-CoV-2. It is possible that these pangolins died during the acute stage of disease, before the appearance of antibodies. Histological examinations of tissues from four betacoronavirus-positive Malayan pangolins revealed diffuse alveolar damage of varying severity in the lung, compared with lung tissue from a betacoronavirus-negative Malayan pangolin. In one case, alveoli were filled with desquamated epithelial cells and some macrophages with haemosiderin pigments, with considerably reduced alveolar space, leading to the consolidation of the lung. In other cases, similar changes

---

[1]Center for Emerging and Zoonotic Diseases, College of Veterinary Medicine, South China Agricultural University, Guangzhou, China. [2]Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou, China. [3]Guangzhou Zoo & Guangzhou Wildlife Research Center, Guangzhou, China. [4]Guangdong Provincial Wildlife Rescue Center, Guangzhou, China. [5]State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China. [6]Zhaoqing Branch Center of Guangdong Laboratory for Lingnan Modern Agricultural Science and Technology, Zhaoqing, China. [7]These authors contributed equally: Kangpeng Xiao, Junqiong Zhai. ✉e-mail: guangzhouchenwu@sina.com; lxiao@scau.edu.cn; shenyy@scau.edu.cn

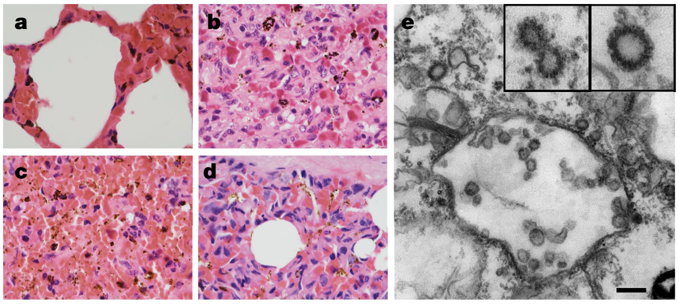


Fig. 1 | Pathological changes in the lungs of pangolins that are potentially induced by pangolin-CoV. a–d, Histological changes in the lung tissues are compared between a virus-negative Malayan pangolin (**a**) and three Malayan pangolins naturally infected with pangolin-CoV (**b**–**d**) (original magnification ×1,000). Proliferation and desquamation of alveolar epithelial cells and haemosiderin pigments are seen in tissues from all three infected pangolins and severe capillary congestion is seen in one of them (**c**). **e**, Viral particles are seen in double-membrane vesicles in the transmission electron microscopy image taken from Vero E6 cell culture inoculated with supernatant of homogenized lung tissue from one pangolin, with morphology indicative of coronavirus (inserts at the top right corner of **e**). Scale bar, 200 nm.

were more focal (Fig. 1, Extended Data Fig. 1). The severe case also had exudate with red blood cells and necrotic cell debris in bronchioles and bronchi. Focal mononuclear-cell infiltration was seen in the bronchioles and bronchi in two of the cases, and haemorrhage was seen in the bronchioles and small bronchi in one case (Extended Data Figs. 1–3). Hyaline membrane and syncytia were not detected in the alveoli of the four cases we examined.

To isolate the virus, supernatant from homogenized lung tissue from one dead Malayan pangolin was inoculated into Vero E6 cells. Obvious cytopathogenic effects were observed in cells after a 72-h incubation. Viral particles were detected by transmission electron microscopy: most of these particles were inside double-membrane vesicles, with a few outside of them. They showed typical coronavirus morphology (Fig. 1e). RT–PCR targeting the spike (*S*) and *RdRp* genes produced the expected PCR products: these PCR products had approximately 84.5% and 92.2% nucleotide sequence identity, respectively, to the partial *S* and *RdRp* genes of SARS-CoV-2.

Illumina RNA sequencing was used to identify viruses in the lung from nine pangolins. Mapping sequence data to the reference SARS-CoV-2 WHCV genome identified coronavirus sequence reads in seven samples (Extended Data Table 3). For one sample, higher genome coverage was obtained by remapping the total reads to the reference genome (Extended Data Fig. 4). We obtained the completed coronavirus genome (29,825 bp)—which we designated pangolin-CoV—using the assembled contigs, short sequence reads and targeted PCR analysis. The full *S* gene was sequenced in six PCR-positive samples, which revealed the presence of only four nucleotide differences in the sequence alignment among these samples (Extended Data Fig. 5); this indicates that only one type of coronavirus was present in the batch of study samples. The predicted *S*, *E*, *M* and *N* genes of pangolin-CoV are 3,798, 228, 669 and 1,260 bp, respectively, in length and the proteins they encode share 90.7%, 100%, 98.6% and 97.8% amino acid identity to the equivalent proteins of SARS-CoV-2 (Table 1).

In a Simplot analysis of whole-genome sequences, we found that pangolin-CoV was highly similar to SARS-CoV-2 and RaTG13, with sequence identity between 80 and 98% (except for the *S* gene) (Fig. 2). Further comparative analysis of the *S* gene sequences suggests that there were recombination events among some of the SARS-related coronaviruses that we analysed. In the region of nucleotides 1–914, pangolin-CoV is more similar to the bat SARS-related coronaviruses ZXC21 and ZC45, whereas in the remaining part of the gene pangolin-CoV is more similar to SARS-CoV-2 and RaTG13 (Fig. 2). In particular, the receptor-binding domain (RBD) of the S protein of pangolin-CoV has only one amino acid difference with SARS-CoV-2. Overall, these data indicate that SARS-CoV-2 might have originated from the recombination of a virus similar to pangolin-CoV and a virus similar to RaTG13 (Fig. 2). To further support this conclusion, we assessed the evolutionary relationships among betacoronaviruses in the full genome, the *RdRp* and *S* genes, and in different regions of the *S* gene (Fig. 2c, Extended Data Fig. 6). The topologies mostly showed the clustering of pangolin-CoV with SARS-CoV-2 and RaTG13; SARS-CoV-2 and RaTG13 form a subclade within this cluster (Fig. 2c). However, pangolin-CoV and SARS-CoV-2 grouped together in the phylogenetic analysis of the RBD. Conflicts in cluster formation among phylogenetic analyses of different regions of the genome serve as a strong indication of genetic recombination, as has previously been seen for SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV)[6,7].

As the S proteins of both SARS-CoV and SARS-CoV-2 have previously been shown to specifically recognize angiotensin-converting enzyme 2 (ACE2) during the entry of host cells[2,8], we conducted molecular binding simulations of the interaction of the S proteins of the four closely related SARS-related coronaviruses with ACE2 proteins from humans, civets and pangolins. As expected, the RBD of SARS-CoV binds efficiently to ACE2 from humans and civets in the molecular binding simulation. In addition, this RBD appears to be capable of binding ACE2 of pangolins. By contrast, the S proteins of SARS-CoV-2 and pangolin-CoV can potentially recognize only the ACE2 of humans and pangolins (Extended Data Fig. 7).

SARS-CoV-2 is one of three known zoonotic coronaviruses (the others are SARS-CoV and MERS-CoV) that infect the lower respiratory tract and cause severe respiratory syndromes in humans[7,9]. Thus far, SARS-CoV-2 has been more contagious, but less deadly, than SARS-CoV[10]: the total number of human infections by SARS-CoV-2 far exceeds those of SARS-CoV[11]. Epidemiological investigations of the SARS-CoV-2 outbreak have shown that some of the initial patients were associated with the Huanan seafood market, where live wildlife was also sold[10]. No animals thus far have been implicated as carriers of the virus. SARS-CoV-2 forms a cluster with SARS-CoV and bat SARS-related coronaviruses (Fig. 2c). In addition, a bat coronavirus (RaTG13) has about 96% sequence identity to SARS-CoV-2 at the whole-genome level[2]. Therefore, it is reasonable to assume that bats are the native host of SARS-CoV-2, as has previously been suggested for SARS-CoV and MERS-CoV[12,13]. The SARS-related coronavirus identified in the present study and the metagenomic

Table 1 | Genomic comparison of pangolin-CoV with SARS-CoV-2, SARS-CoV and bat SARS-related coronaviruses

	S	*E*	*M*	*N*	Full-length genome
WHCV	84.5 (90.7)	99.1 (100)	93.2 (98.6)	96.1 (97.8)	90.1
SARS-CoV GD01	72.2 (77.2)	93.5 (93.5)	85.8 (90.0)	87.5 (90.0)	81.6
RaTG13	88.5 (89.8)	99.6 (100)	93.6 (99.1)	94.0 (96.7)	88.9
ZC45	83.1 (86.1)	98.7 (100)	94.2 (99.6)	88.9 (93.3)	88.0
ZXC21	81.1 (85.4)	98.7 (100)	94.2 (99.6)	88.9 (93.3)	88.4

Numbers represent the percentage of nucleotides shared; numbers in parentheses represent percentage of amino acids shared.

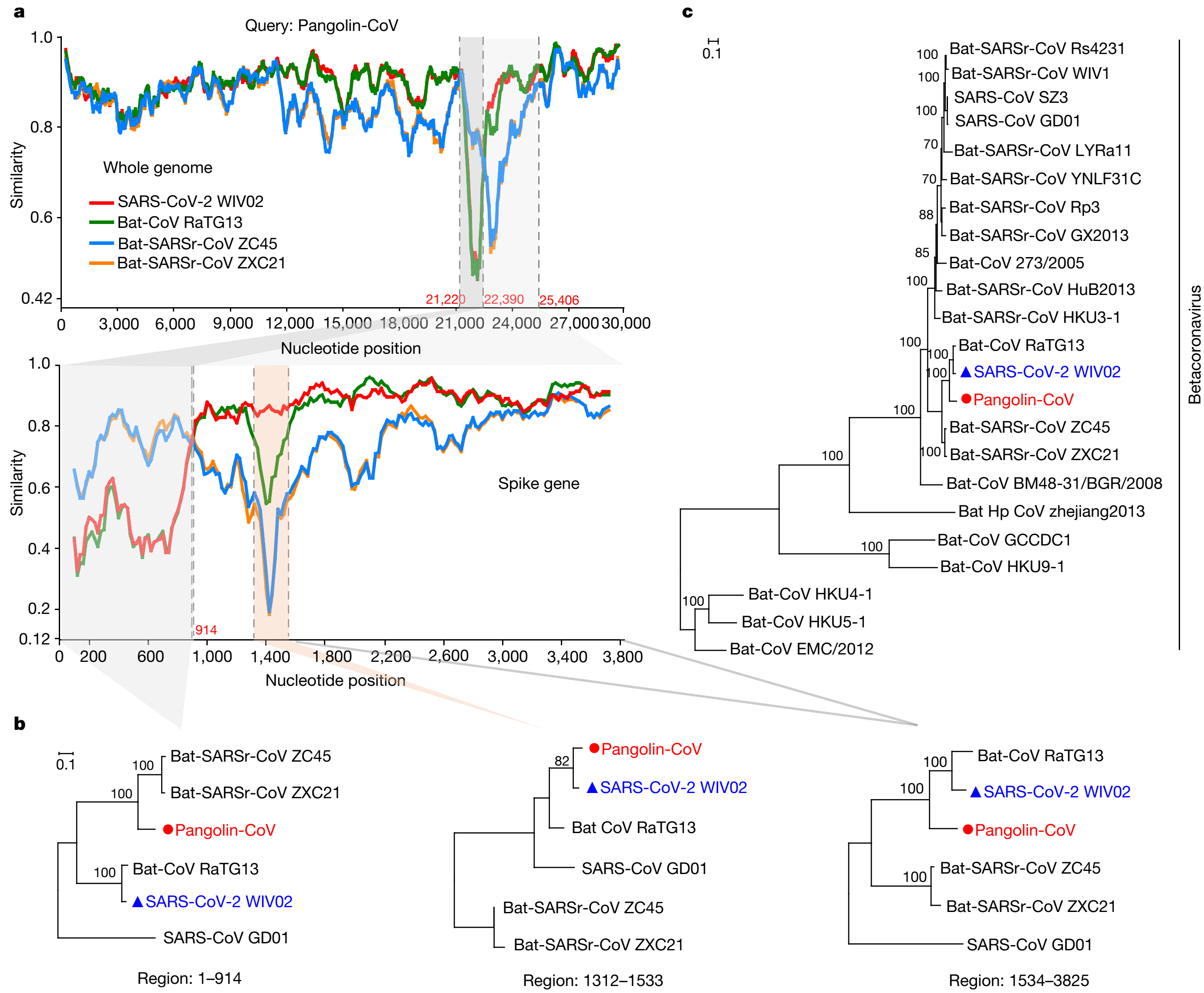


Fig. 2 | Genome characterization of pangolin-CoV. a, Similarity plot of the full-length genomes and *S* gene sequences of pangolin-CoV against sequences of SARS-CoV-2 strain WIV02, as well as RaTG13, ZC45 and ZXC21. Although pangolin-CoV has a high sequence identity to SARS-CoV-2 and RaTG13 in most regions of the *S* gene, it is more similar to ZXC21 and ZC45 at the 5′ end. SARS-rCoV, SARS-related coronavirus. Parameters for the similarity plots are: window, 500 bp; step, 50 bp; gap strip, on; Kimura (2 parameter); *T*/*t* 2.0. **b**, Because of the presence of genetic recombination, there is discrepancy in cluster formation among the outcomes of phylogenetic analyses of different regions of the *S* gene. **c**, Phylogeny of coronaviruses closely related to SARS-CoV-2, based on full genome sequences. The phylogenetic tree was constructed using RAxML with the substitution model GTRGAMMAI and 1,000 bootstrap replicates. Numbers (>70) above or below branches are percentage bootstrap values for the associated nodes. The scale bar represents the number of substitutions per site. Red circles indicate the pangolin coronavirus sequences generated in this study, and blue triangles indicate SARS-CoV-2 sequences from humans.

assemblies of viral sequences from Malayan pangolins[14] is genetically related to SARS-CoV-2, but is unlikely to be directly linked to the current outbreak because of its substantial sequence differences from SARS-CoV-2. However, a virus related to pangolin-CoV appears to have donated the RBD to SARS-CoV-2. SARS-related coronavirus sequences have previously been detected in dead Malayan pangolins[15]. These sequences appear to be from the same virus (pangolin-CoV) that we identified in the present study, as judged from their sequence similarity. Here we provide evidence for the potential for pangolins to act as the zoonotic reservoir of SARS-CoV-2-like coronaviruses. However, the pangolins we studied here showed clinical signs of disease. In general, a natural reservoir host does not show severe disease, whereas an intermediate host may have clinical signs of infection[16]. Although a SARS-CoV-2-like coronavirus was detected in the lungs of these pangolins, a direct association between the clinical signs or pathology and active virus replication is not available as we lack evidence from immunohistochemistry or in situ hybridization experiments. The experimental infection of healthy pangolins with pangolin-CoV would provide more definitive answers; however, as pangolins are protected it is difficult to carry out such experiments. Further studies are needed to confirm the role of pangolins in the transmission of SARS-related coronaviruses.

As the RBD of pangolin-CoV is nearly identical to that of SARS-CoV-2, the virus in pangolins presents a potential future threat to public health. Pangolins and bats are both nocturnal animals, eat insects and share overlapping ecological niches[17,18], which make pangolins an ideal intermediate host for some SARS-related coronaviruses. Therefore, more systematic and long-term monitoring of SARS-related coronaviruses in pangolins and related animals should be implemented to identify the potential animal source of SARS-CoV-2 in the current outbreak.

Our findings support the call for stronger enforcement of regulations against the illegal trade in pangolins. Owing to the demand for their meat as a delicacy and their scales for use in traditional medicine in China, the illegal smuggling of pangolins from Southeast Asia to China is widespread[18]. International co-operation in the implementation of stricter regulations against illegal wildlife trade and consumption of game meat should be encouraged, as this will increase the protection of endangered animals and help to prevent future outbreaks of diseases caused by SARS-related coronaviruses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2313-x.

1. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med*. **382**, 727–733 (2020).
2. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
3. Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol*. **17**, 181–192 (2019).
4. Banerjee, A., Kulcsar, K., Misra, V., Frieman, M. & Mossman, K. Bats and coronaviruses. *Viruses* **11**, 41 (2019).
5. Masters, P. S. The molecular biology of coronaviruses. *Adv. Virus Res*. **66**, 193–292 (2006).
6. Hu, B. et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog*. **13**, e1006698 (2017).
7. Sabir, J. S. M. et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* **351**, 81–84 (2016).
8. Li, W. et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
9. Song, Z. et al. From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses* **11**, E59 (2019).
10. Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med*. **382**, 1199–1207 (2020).
11. WHO. *Coronavirus Disease 2019 (COVID-19) Situation Report – 79*, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200408-sitrep-79-covid-19.pdf?sfvrsn=4796b143_4 (WHO, 2020).
12. Ge, X. Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
13. Li, W. et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
14. Lam, T. T.-Y. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* https://doi.org/10.1038/s41586-020-2169-0 (2020).
15. Liu, P., Chen, W. & Chen, J. P. Viral metagenomics revealed sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses* **11**, 979 (2019).
16. Wu, D. et al. Civets are equally susceptible to experimental infection by two different severe acute respiratory syndrome coronavirus isolates. *J. Virol*. **79**, 2620–2625 (2005).
17. Nowak, R. M. *Walker's Bats of the World* (Johns Hopkins Univ. Press, 1994).
18. Shepherd, C. R. Overview of the pangolin trade in southeast Asia. In *Proc. Workshop on Trade and Conservation of Pangolins Native to South and Southeast Asia: 30 June-2 July 2008, Singapore Zoo* (eds Pantel, S. & Chin, S. Y.) 6–12 (TRAFFIC Southeast Asia, 2009).

Methods

No statistical methods were used to predetermine sample size.

Metagenomic analysis and viral genome assembly

We collected viromic, metagenomic and transcriptomic data of different mammals and birds in public databases—including NCBI Sequence Read Archive (SRA) and European Nucleotide Archive (ENA)—for searching potential coronavirus sequences. The raw reads from the public databases and some in-house metagenomic datasets were trimmed using fastp (v.0.19.7)[19] to remove adaptor and low-quality sequences. The clean reads were mapped to the SARS-CoV-2 reference sequence (MN908947) using BWA-MEM (v.0.7.17)[20] with >30% matches. The mapped reads were collected for downstream analyses. Contigs were de novo-assembled using Megahit (v.1.0.3)[21] and identified as related to SARS-CoV-2 using BLASTn with E-values $< 1 \times 10^{-5}$ and sequence identity >90%.

Samples

Pangolins used in the study were confiscated by Customs and Department of Forestry of Guangdong Province in March and August 2019. They included four Chinese pangolins (*M. pentadactyla*) and 25 Malayan pangolins (*M. javanica*). The first transport confiscated contained 21 Malayan pangolins, and the second transport contained 4 Malayan pangolins and 4 Chinese pangolins. These pangolins were sent to the wildlife rescue centre, and were mostly inactive and crying, and eventually died in custody despite exhaustive rescue efforts. Tissue samples were taken from the lung of pangolins that had just died for histological and virological examinations.

Pathological examinations

Histological examinations were performed on lung tissues from five Malayan pangolins. In brief, the tissues collected were cut into small pieces and fixed in 10% buffered formalin for 24 h. They were washed free of formalin, dehydrated in ascending grades of ethanol and cleared with chloroform, and then embedded with molten paraffin wax in a template. The tissue blocks were sectioned with a microtome. The sections were transferred onto grease-free glass slides, deparaffinized and rehydrated through descending grades of ethanol and distilled water. They were stained with a haematoxylin and eosin staining kit (Baso Diagnostics, Wuhan Servicebio Technology). Finally, the stained slides were mounted with coverslips and examined under an Olympus BX53 equipped with an Olympus PM-C 35 camera.

Virus isolation and RT–PCR analysis

Lung tissue extract from pangolins was inoculated into Vero E6 cells for virus isolation. The cell line was tested free of mycoplasma contamination using LookOut Mycoplasma PCR Detection Kit (SIGMA), and was authenticated by microscopic morphologic evaluation. Cultured cell monolayers were maintained in Dulbecco's Modified Eagle Medium (DMEM) and Ham's F-12. The inoculum was prepared by grinding the lung tissue in liquid nitrogen, diluting it 1:2 with DMEM, filtering it through a 0.45-μm filter (Merck Millipore), and treating it with 16 μg/ml trypsin solution. After incubation at 37 °C for 1 h, the inoculum was removed from the culture and replaced with fresh culture medium. The cells were incubated at 37 °C and observed daily for cytopathic effects.

Viral RNA was extracted from the lung tissue using the QIAamp Viral RNA Mini kit (Qiagen) following the manufacturer-recommended procedures, and examined for coronavirus by RT–PCR using a pair of primers (F: 5′-TGGCWTATAGGTTYAATGGYATTGGAG-3′, R: 5′-CCGTCGATTGTGTGWATTTGSACAT-3′) designed to amplify the *S* gene of betacoronavirus.

Transmission electron microscopy

Cell cultures that showed cytopathic effects were examined for the viral particles using transmission electron microscopy. Cells were collected from the culture by centrifugation at 1,000*g* for 10 min, and fixed initially with 2.5% glutaraldehyde solution at 4 °C for 4 h, and again with 1% osmium tetroxide. They were dehydrated with graded ethanol and embedded with PON812 resin. Sections (80 nm in thickness) were cut from the resin block and stained with uranyl acetate and lead citrate sequentially. The negative stained grids and ultrathin sections were observed under a HT7800 transmission electron microscope (Hitachi).

Serological test

Plasma samples from eight Malayan pangolins were tested for anti-SARS-CoV-2 antibodies using a double-antigen ELISA kit for the detection of antibodies against SARS-CoV-2 by Hotgen, following manufacturer-recommended procedures. The assay was designed for the detection of both IgG and IgM antibodies against SARS-CoV-2 in humans and animals, and marketed as supplementary diagnostic tool for COVID-19. It uses the capture of antibodies against SARS-CoV-2 by the S1 antigen precoated on ELISA plates, and the detection of the antibodies through the use of horseradish peroxidase-conjugated RBD. Both the S1 antigen and RBD fragment were expressed in eukaryotic cells. Data generated by the test developer have shown a 95% detection rate in the analysis of sera from over 200 patients with COVID-19s. The assay has an inter-test variation of ≤15%, and no cross-reactivities with sera or plasma from patients positive for SARS-CoV, common and avian influenza viruses, mycoplasma and chlamydia. Fifty microlitres of plasma was analysed in duplicate, together with two negative controls and one positive control. The reaction was read on a Synergy HTX Multi-Mode Microplate Reader (BioTek) at 450/630 nm, with optical density (OD) values being calculated. The cut-off OD value for positivity was 0.105 + mean OD from the negative controls, and the cut-off value for OD for the positive control was set at ≥ 0.5. Positive samples were tested again with serial-diluted plasma.

Metagenomic sequencing

The lung tissue was homogenized by vortex with silica beads in 1 ml of phosphate-buffered saline. The homogenate was centrifuged at 10,000*g* for 5 min, with the supernatant being filtered through a 0.45-μm filter (Merck Millipore) to remove large particles. The filtrate or virus culture supernatant was used in RNA extraction with the QIAamp Viral RNA Mini kit. cDNA was synthesized from the extracted RNA using PrimeScriptScript II reverse transcriptase (Takara) and random primers, and amplified using Klenow Fragment (New England Biolabs). Sequencing libraries were prepared with NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs), and sequenced paired-end (150-bp) on an Illumina NovaSeq 6000. Specific PCR assays were used to fill genome sequence gaps, using primers designed based on sequences flanking the gap.

Phylogenetic analysis

Multiple sequence alignments of all sequence data were constructed using MAFFT v.7.221[22]. The phylogenetic relationship of the viral sequences was assessed using RAxML v.8.0.14[23]. The best-fit evolutionary model for the sequences in each dataset was identified using ModelTest[24]. Potential recombination events and the location of possible breakpoints in betacoronavirus genomes were detected using Simplot (version 3.5.1)[25] and RDP 4.99[26].

Molecular simulation of interactions between RBD and ACE2

The interaction between the RBD of the S protein of SARS-related coronavirus and the ACE2 of humans, civets, and pangolins was examined using molecular dynamic simulation. The crystal structure of SARS-CoV RBD domain binding to human ACE2 protein complex was downloaded from Protein Data Bank (PDB code 2AJF[27]). The structures of the complexes formed by ACE2 of civets or pangolins and the RBD of SARS-CoV-2, RaTG13 and pangolin-CoV were made using the MODELLER program[28], and superimposed with the template (PDB code 2AJF).

The sequence identity of SARS-CoV RBD (PDB code 6ACD) to the RBD of SARS-CoV-2, RaTG13 and pangolin-CoV was 76.5%, 76.8% and 74.2%, respectively, and the sequence identity of the human ACE2 protein to that of pangolins and civets was 85.4% and 86.9%, respectively.

The molecular dynamic simulations of RBD–ACE2 complexes were carried out using the AMBER 18 suite[29] and ff14SB force field[30]. After two-stage minimization, NVT and NPT-MD, a 30-ns production molecular dynamics simulation was applied, with the time step being set to 2 fs and coordinate trajectories being saved every 3 ps. The MM-GBSA[31] approach was used to calculate the binding free energy of each ACE2 protein to the RBD of the S protein, using the python script MMPBSA.py[32] in the build-in procedure of AMBER 18 suite. The last 300 frames of all simulations were extracted to calculate the binding free energy that excludes the contributions of disulfide bond.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequence reads generated in this study are available in the NCBI SRA database under the BioProject accession PRJNA607174. The complete genome sequence of pangolin-CoV has been deposited in GISAID with the accession number EPI_ISL_410721.

19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
21. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
22. Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899–1900 (2010).
23. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
24. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).
25. Lole, K. S. et al. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73**, 152–160 (1999).
26. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).
27. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868 (2005).
28. Webb, B. & Sali, A. Comparative protein structure modeling Using MODELLER. *Curr. Protoc. Bioinformatics* **47**, 5.6.1–5.6.32 (2014).
29. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
30. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
31. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
32. Miller, B. R. III et al. MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).

Acknowledgements This work was supported by the National Natural Science Foundation of China (grant no. 31822056 and 31820103014), National Key R & D Program of China (2017YFD0500404), Fund for the Key Program and Creative Research Group of the Department of Education of Guangdong province (2019KZDXM004 and 2019KCXTD001), Guangdong Science and Technology Innovation Leading Talent Program (2019TX05N098), the Major Program of Guangdong Basic and Applied Research, the 111 Project (D20008), Chinese Academy of Engineering (2020-KYGG-04-01), Department of Science and Technology of Guangdong Province (2020B1111320002) and Department of Agriculture of Guangdong province.

Author contributions Y.S., L.X. and W.C. conceived the study; J.-J.Z., F.-H.H., Y.-J.W., S.-M.P., M.H., W.-J.X., Q.-H.C. and W.C. collected samples; J.Z., N.Z., X.Z., N.L., Y.G., X.L., X.S., Zhipeng Zhang, F.S. and W.H. performed virus isolation and sequencing; K.X., Y.F., Y.L., Ziding Zhang and Y.S. contributed to the analysis; Y.S. and L.X. wrote the manuscript; Y.F. and R.-A.C. edited the manuscript.

Competing interests The authors declare no competing interests.

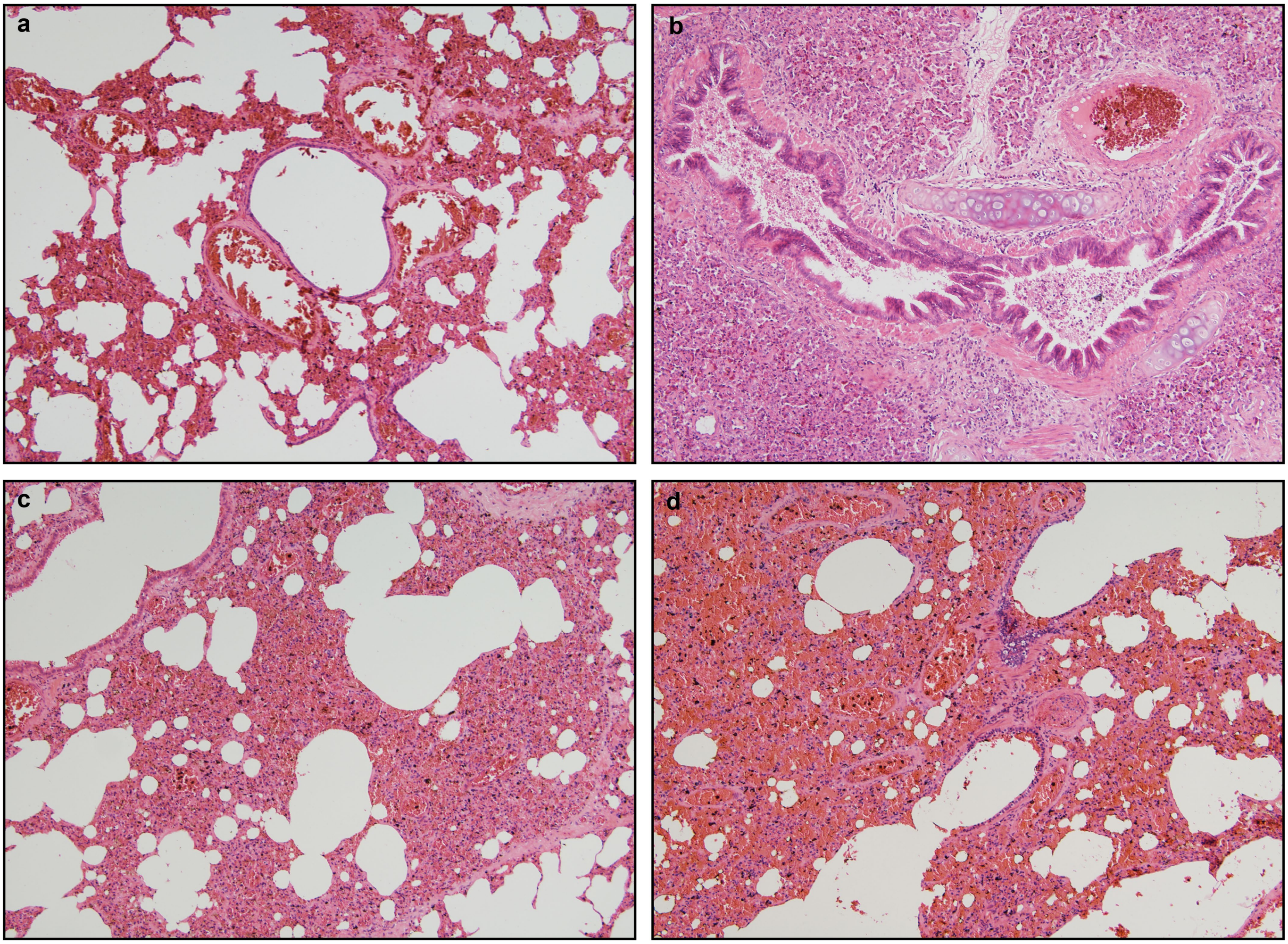
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2313-x.

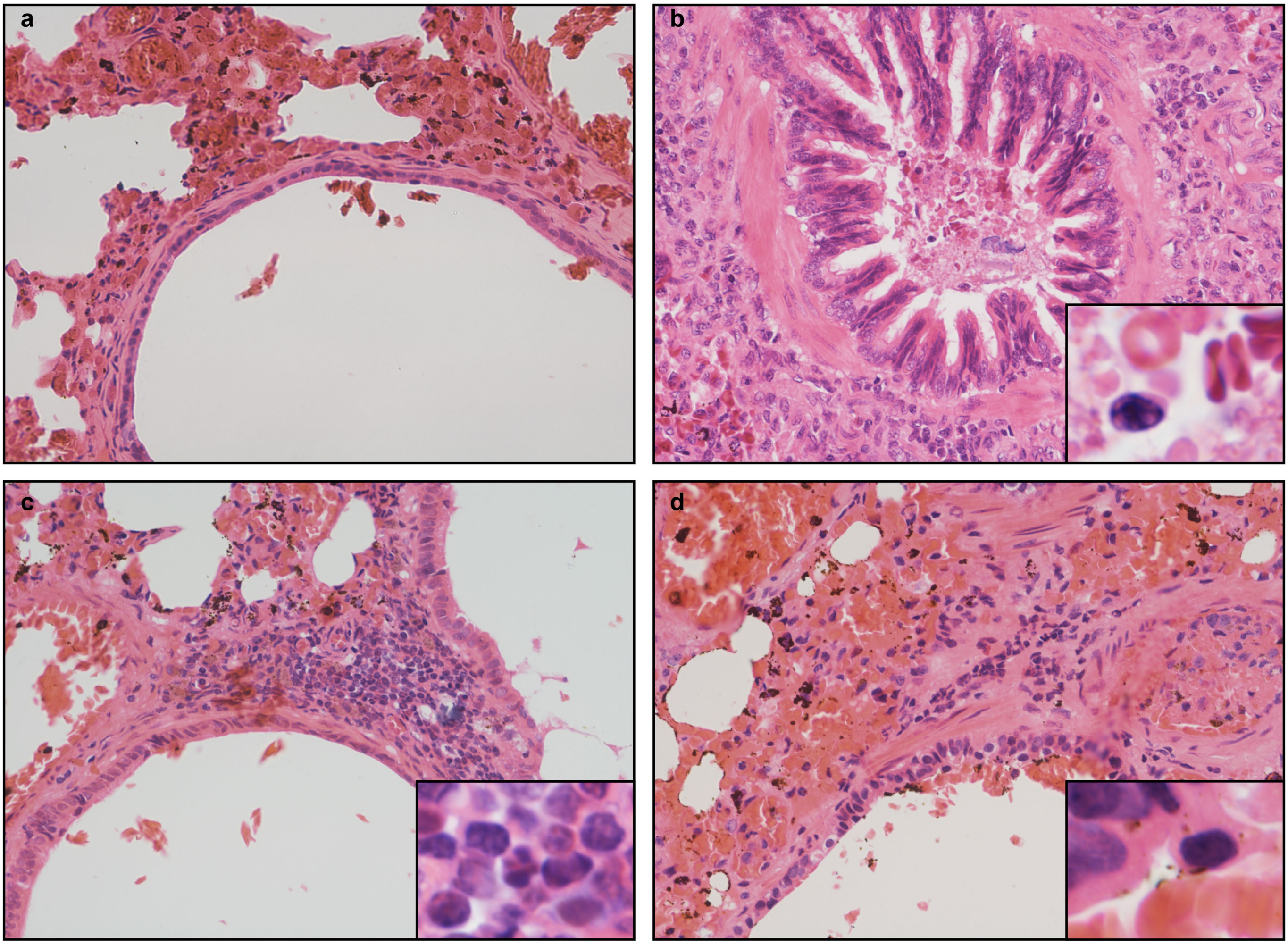
Correspondence and requests for materials should be addressed to W.C., L.X. or Y.S.

Peer review information *Nature* thanks Andrew Ward and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

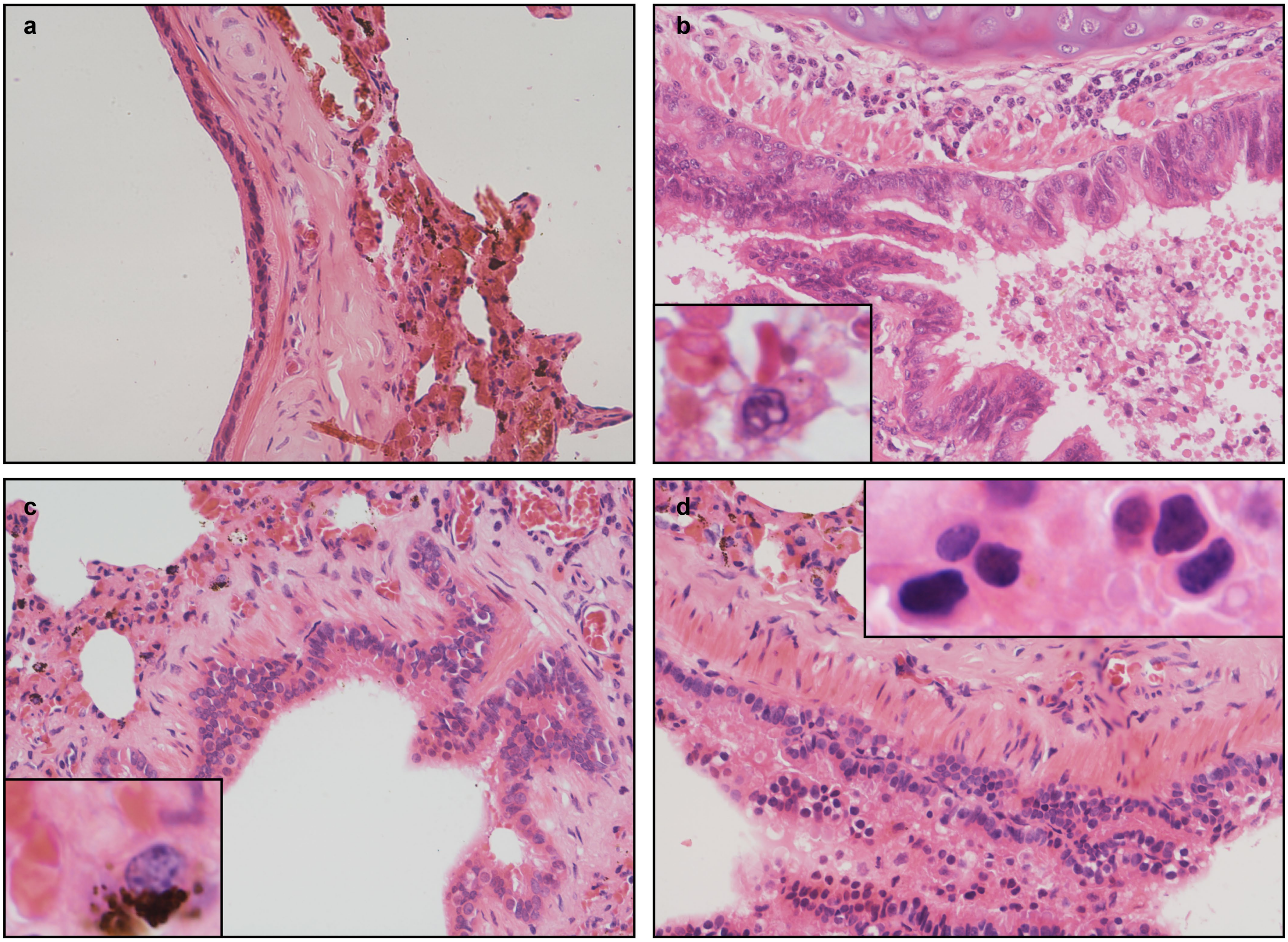
Reprints and permissions information is available at http://www.nature.com/reprints.



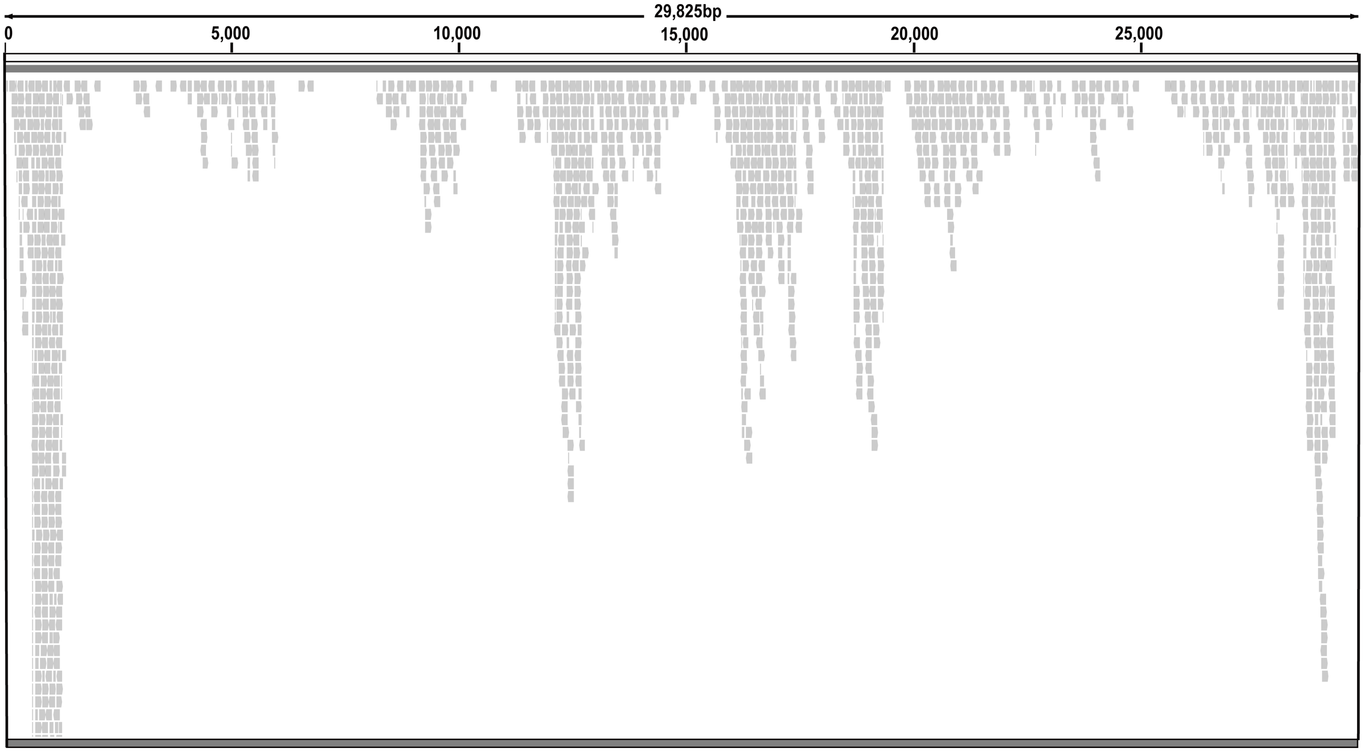
Extended Data Fig. 1 | Pathological changes in the lungs of pangolins. **a**–**d**, Lungs of three Malayan pangolins naturally infected with pangolin-CoV (**b**–**d**, original amplification × 100) in comparison with the lung from a virus-negative Malayan pangolin (**a**). Different degrees of consolidation are seen in the lung tissues from three infected pangolins (**b**–**d**). Exudate is seen in the bronchi of one infected pangolin (**b**). Severe congestion is seen in the lung of one pangolin (**d**).



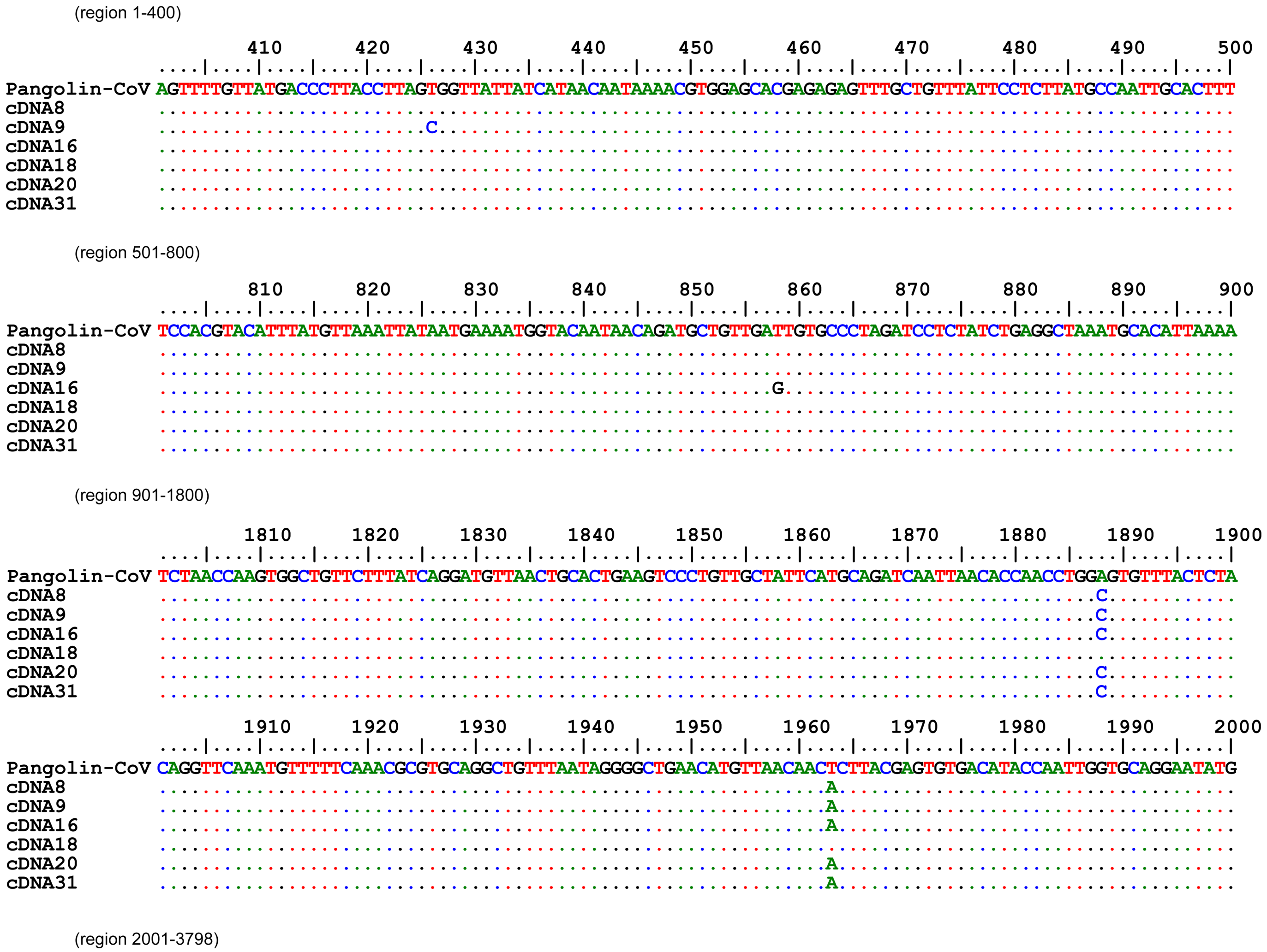
Extended Data Fig. 2 | Pathological changes in the bronchiole of pangolins. **a**–**d**, Three Malayan pangolins positive for pangolin-CoV (**b**–**d**, original amplification × 100) in comparison with a virus-negative Malayan pangolin (**a**). Red blood cells are seen in the bronchioles of two infected pangolins (**b**, **d**). Mononuclear cell infiltration is seen in the bronchiole wall of one infected pangolin (**c**). Severe congestion is seen in the alveolar tissue (in close proximity to the bronchiole) of one pangolin (**d**). The respiratory epithelium in the bronchioles is intact.



Extended Data Fig. 3 | Pathological changes in the bronchus of pangolins. **a**–**d**, Three Malayan pangolins positive for pangolin-CoV (**b**–**d**, original amplification × 100) in comparison with a virus-negative Malayan pangolin (**a**). Exudate with red blood cells is seen in the bronchus of one infected pangolin (**b**). Macrophages with haemosiderin pigments and mononuclear cell infiltration are seen in the bronchus wall of two infected pangolins (**c** and **d**, respectively). The respiratory epithelium in the bronchi is intact.

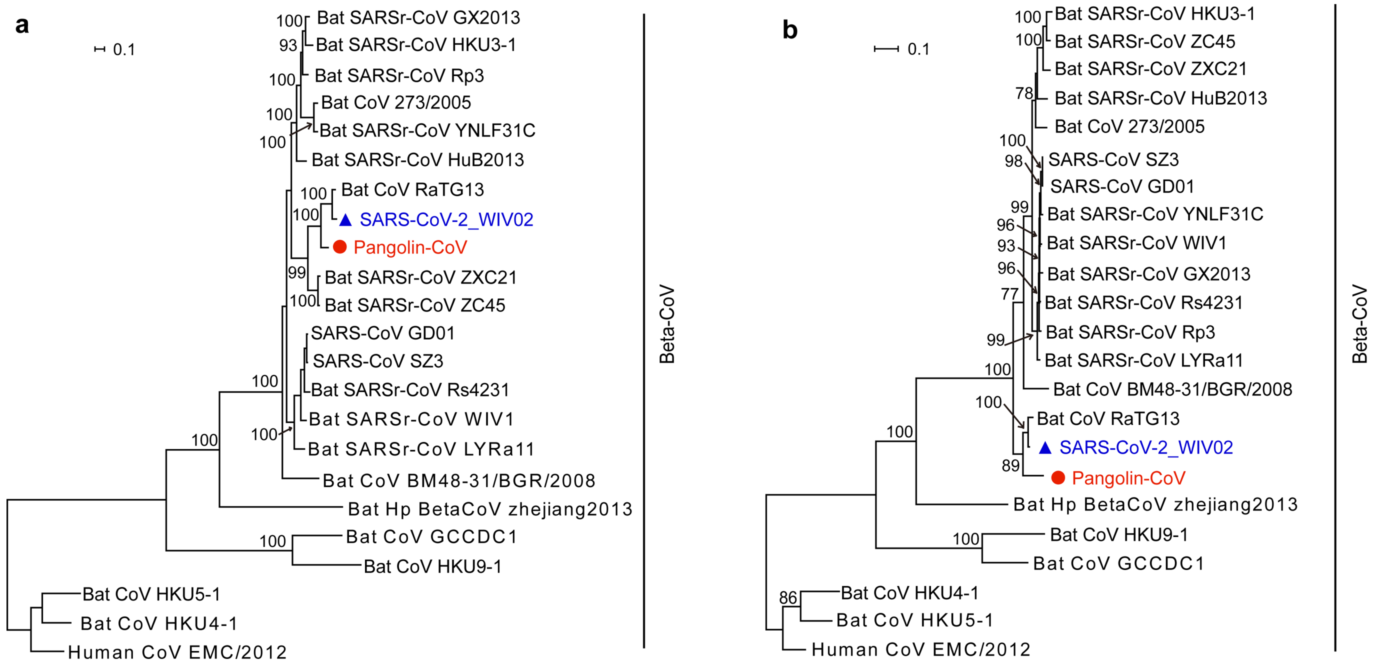


Extended Data Fig. 4 | Mapping raw reads from pangolin lung to pangolin-CoV genome. Results of the mapping of raw reads from the high-throughput sequencing of the pangolin lung tissue to the assembled pangolin-CoV genome.

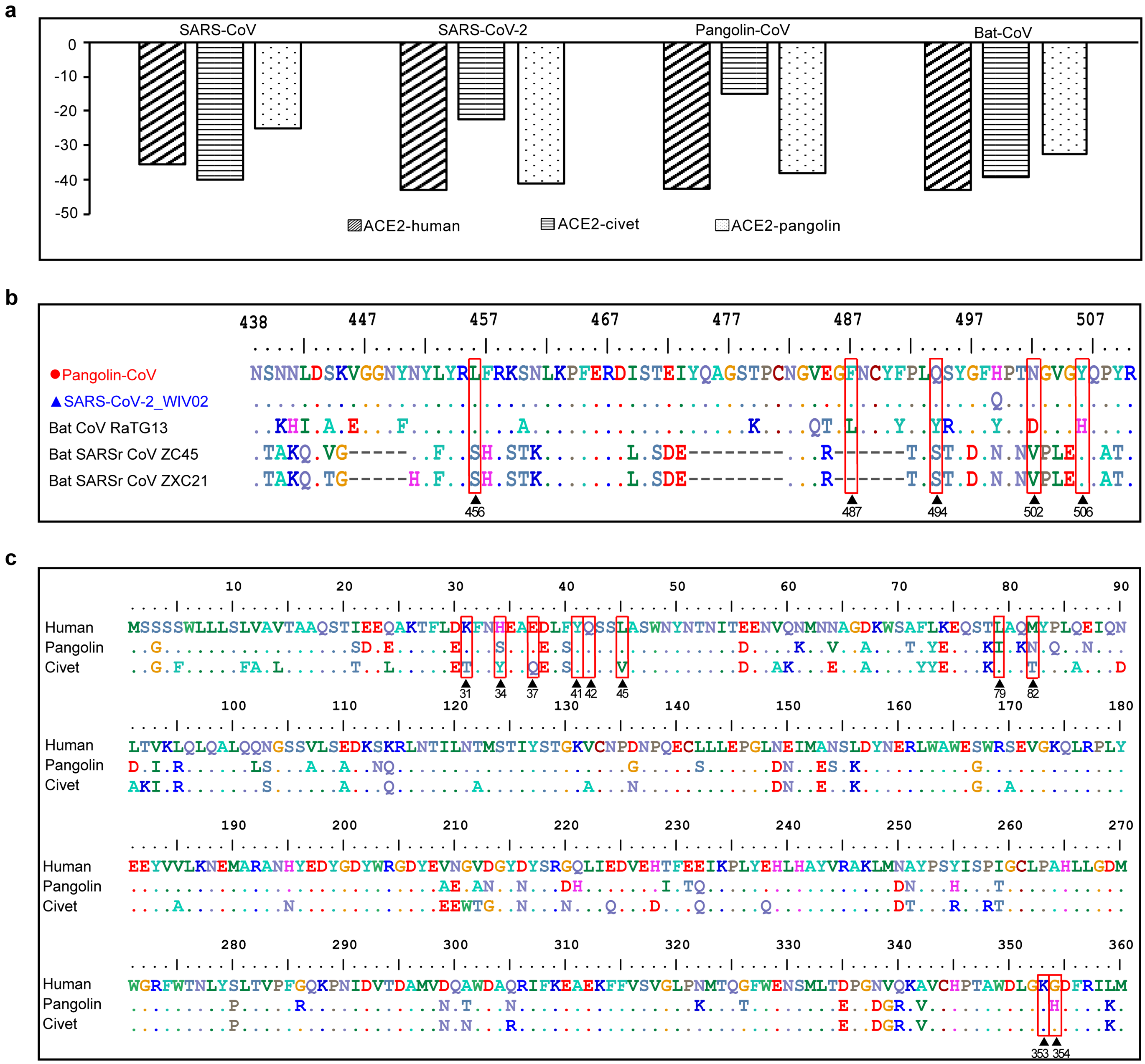


```
(region 1-400)

                    410       420       430       440       450       460       470       480       490       500
             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Pangolin-CoV AGTTTTGTTATGACCCTTACCTTAGTGGTTATTATCATAACAATAAAACGTGGAGCACGAGAGAGTTTGCTGTTTATTCCTCTTATGCCAATTGCACTTT
cDNA8        ....................................................................................................
cDNA9        .........................C..........................................................................
cDNA16       ....................................................................................................
cDNA18       ....................................................................................................
cDNA20       ....................................................................................................
cDNA31       ....................................................................................................

(region 501-800)

                    810       820       830       840       850       860       870       880       890       900
             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Pangolin-CoV TCCACGTACATTTATGTTAAATTATAATGAAAATGGTACAATAACAGATGCTGTTGATTGTGCCCTAGATCCTCTATCTGAGGCTAAATGCACATTAAAA
cDNA8        ....................................................................................................
cDNA9        ....................................................................................................
cDNA16       .........................................................G..........................................
cDNA18       ....................................................................................................
cDNA20       ....................................................................................................
cDNA31       ....................................................................................................

(region 901-1800)

                   1810      1820      1830      1840      1850      1860      1870      1880      1890      1900
             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Pangolin-CoV TCTAACCAAGTGGCTGTTCTTTATCAGGATGTTAACTGCACTGAAGTCCCTGTTGCTATTCATGCAGATCAATTAACACCAACCTGGAGTGTTTACTCTA
cDNA8        .......................................................................................C............
cDNA9        .......................................................................................C............
cDNA16       .......................................................................................C............
cDNA18       ....................................................................................................
cDNA20       .......................................................................................C............
cDNA31       .......................................................................................C............

                   1910      1920      1930      1940      1950      1960      1970      1980      1990      2000
             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Pangolin-CoV CAGGTTCAAATGTTTTTCAAACGCGTGCAGGCTGTTTAATAGGGGCTGAACATGTTAACAACTCTTACGAGTGTGACATACCAATTGGTGCAGGAATATG
cDNA8        ..............................................................A.....................................
cDNA9        ..............................................................A.....................................
cDNA16       ..............................................................A.....................................
cDNA18       ....................................................................................................
cDNA20       ..............................................................A.....................................
cDNA31       ..............................................................A.....................................

(region 2001-3798)
```

Extended Data Fig. 5 | Sequence polymorphism among nucleotide sequences of the full *S* gene among virus-positive lung samples from six pangolins. Dots denote nucleotide identity to the reference sequence.



Extended Data Fig. 6 | Phylogeny of coronaviruses closely related to SARS-CoV-2. a, Based on nucleotide sequences of the *S* gene. **b**, Based on nucleotide sequences of the *RdRp* gene. The phylogenetic trees were constructed by RAxML with the substitution model GTRGAMMAI and 1,000 bootstrap replicates. Numbers (>70) above or below branches are percentage bootstrap values for the associated nodes. The scale bar represents the number of substitutions per site. Red circles indicate the pangolin coronavirus sequences generated in this study, and blue triangles indicate SARS-CoV-2 sequences from humans.



Extended Data Fig. 7 | Molecular binding simulations of the interaction of the S proteins of four closely related SARS-related coronaviruses with the ACE2 proteins of humans, civets and pangolins. **a**, Free energy (kcal mol^{-1}) for the binding of the RBD of S proteins of four SARS-related coronaviruses to the ACE2 of potential hosts. **b**, Alignment of the RBD sequences (key amino acids involved in interactions with ACE2 are boxed) of the S proteins from several genetically related SARS-related coronaviruses. **c**, Alignment of partial ACE2 amino acid sequences (key amino acids involved in interactions with RBD are marked with arrowheads) from humans, pangolins and civets at their interface with the RBD of S proteins.

Extended Data Table 1 | Results of Blast search of SARS-related coronavirus sequences in available mammalian and avian viromic, metagenomic and transcriptomic data using the SARS-CoV-2 sequence (GenBank accession number MN908947)

Contig name	Sequence identity (%)	Length (bp)	E-value	Data source
con_9	98.623	363	0	PRJNA573298
con_15	97.303	519	0	PRJNA573298
con_2	96.97	203	2.41E-44	PRJNA573298
con_19	96.855	477	0	PRJNA573298
con_10	96.562	349	4.64E-168	PRJNA573298
con_1	95.96	203	1.12E-42	PRJNA573298
con_13	95.918	373	7.65E-42	PRJNA573298
con_11	95.804	402	5.63E-133	PRJNA573298
con_4	94.737	231	1.00E-38	PRJNA573298
con_25	94.737	779	0	PRJNA573298
con_22	94.297	655	5.73E-115	PRJNA573298
con_7	94.098	305	1.18E-133	PRJNA573298
con_16	94.041	387	3.08E-170	PRJNA573298
con_21	93.971	680	0	PRJNA573298
con_6	93.96	304	9.13E-130	PRJNA573298
con_24	93.74	610	0	PRJNA573298
con_20	93.343	721	0	PRJNA573298
con_12	93.333	373	4.14E-114	PRJNA573298
con_30	93.321	1048	0	PRJNA573298
con_29	92.892	886	0	PRJNA573298
con_5	92.632	231	2.17E-35	PRJNA573298
con_31	92.495	635	0	PRJNA573298
con_23	92.354	669	0	PRJNA573298
con_32	91.942	1031	0	PRJNA573298
con_34	91.884	1687	0	PRJNA573298
con_27	91.844	846	0	PRJNA573298
con_8	91.776	304	1.99E-121	PRJNA573298
con_3	91.705	218	6.83E-85	PRJNA573298
con_14	91.436	410	5.55E-158	PRJNA573298
con_18	91.429	385	5.23E-153	PRJNA573298
con_33	91.358	1177	7.06E-27	PRJNA573298
con_17	91.02	491	0	PRJNA573298
con_26	90.921	740	0	PRJNA573298
con_28	90.814	840	0	PRJNA573298

Extended Data Table 2 | OD values (at 450 and 630 nm) of ELISA testing of SARS-CoV-2 antibodies in plasma samples from eight pangolins

Repetition / Sample	Repetition 1	Repetition 2	Average
1	2.253	2.088	2.1705
2	0.014	0.012	0.013
3	0.013	0.012	0.0125
4	0.023	0.025	0.024
5	0.01	0.012	0.011
6	0.011	0.011	0.011
7	0.028	0.028	0.028
8	0.053	0.052	0.0525
Negative control	0.01	0.01	0.01

Extended Data Table 3 | Identification of SARS-related coronavirus sequence reads in metagenomes from the lung of pangolins using the SARS-CoV-2 sequence (GenBank accession number MN908947) as the reference

Sample ID	Animal species	Total reads*	No. mapped
M1	Malayan pangolin	107,267,359	496
M2	Malayan pangolin	38,091,846	302
M3	Malayan pangolin	79,477,358	14
M4	Malayan pangolin	32,829,850	1,100
M5	Malayan pangolin	547,302,862	56
M6	Malayan pangolin	232,433,120	10
M8	Malayan pangolin	44,440,374	12
M10	Malayan pangolin	227,801,882	0
Z1	Chinese pangolin	444,573,526	0

* 150-bp paired-end reads

nature research

Corresponding author(s): Yongyi Shen, Lihua Xiao, Wu Chen

Last updated by author(s): Apr 22, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	*Provide a description of all commercial, open source and custom code used to analyse the data in this study, specifying the version used OR state that no software was used.*

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data that support the findings of this study have been deposited in GISAID with the accession numbers EPI_ISL_410721. Raw data of RNAseq are available from the NCBI SRA under the study accession number PRJNA607174.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Article

Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody

https://doi.org/10.1038/s41586-020-2349-y

Received: 6 April 2020

Accepted: 12 May 2020

Dora Pinto[1,9], Young-Jun Park[2,9], Martina Beltramello[1,9], Alexandra C. Walls[2,9], M. Alejandra Tortorici[2,3], Siro Bianchi[1], Stefano Jaconi[1], Katja Culap[1], Fabrizia Zatta[1], Anna De Marco[1], Alessia Peter[1], Barbara Guarino[1], Roberto Spreafico[4], Elisabetta Cameroni[1], James Brett Case[5], Rita E. Chen[5,6], Colin Havenar-Daughton[4], Gyorgy Snell[4], Amalio Telenti[4], Herbert W. Virgin[4], Antonio Lanzavecchia[1,7], Michael S. Diamond[5,6,8], Katja Fink[1], David Veesler[2]✉ & Davide Corti[1]✉

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a newly emerged coronavirus that is responsible for the current pandemic of coronavirus disease 2019 (COVID-19), which has resulted in more than 3.7 million infections and 260,000 deaths as of 6 May 2020[1,2]. Vaccine and therapeutic discovery efforts are paramount to curb the pandemic spread of this zoonotic virus. The SARS-CoV-2 spike (S) glycoprotein promotes entry into host cells and is the main target of neutralizing antibodies. Here we describe several monoclonal antibodies that target the S glycoprotein of SARS-CoV-2, which we identified from memory B cells of an individual who was infected with severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003. One antibody (named S309) potently neutralizes SARS-CoV-2 and SARS-CoV pseudoviruses as well as authentic SARS-CoV-2, by engaging the receptor-binding domain of the S glycoprotein. Using cryo-electron microscopy and binding assays, we show that S309 recognizes an epitope containing a glycan that is conserved within the *Sarbecovirus* subgenus, without competing with receptor attachment. Antibody cocktails that include S309 in combination with other antibodies that we identified further enhanced SARS-CoV-2 neutralization, and may limit the emergence of neutralization-escape mutants. These results pave the way for using S309 and antibody cocktails containing S309 for prophylaxis in individuals at a high risk of exposure or as a post-exposure therapy to limit or treat severe disease.

The entry of coronaviruses into host cells is mediated by the transmembrane S glycoprotein, which forms homotrimers that protrude from the viral surface[3]. The S glycoprotein comprises two functional subunits: S_1 (divided into A, B, C and D domains), which is responsible for binding to host-cell receptors; and S_2, which promotes fusion of the viral and cellular membranes[4,5]. Both SARS-CoV-2 and SARS-CoV belong to the *Sarbecovirus* subgenus and their S glycoproteins share 80% amino acid sequence identity[6]. SARS-CoV-2 S glycoprotein is closely related to the bat SARS-related coronavirus RaTG13 S, with which it shares 97.2% amino acid sequence identity[1]. It has recently been demonstrated that, in humans, angiotensin converting enzyme 2 (ACE2) is a functional receptor for SARS-CoV-2, as also is the case for SARS-CoV[1,6–8]. Domain B of subunit S_1 (S^B) is the receptor-binding domain (RBD) of the S glycoprotein, and binds to ACE2 with high affinity, which possibly contributed to the current rapid transmission of SARS-CoV-2 in humans[6,9] as was previously proposed for SARS-CoV[10].

As the S glycoprotein of coronaviruses mediates entry into host cells, it is the main target of neutralizing antibodies and the focus of efforts to design therapeutic agents and vaccines[3]. The S-glycoprotein trimers are extensively decorated with N-linked glycans that are important for protein folding[11] and modulate accessibility to host proteases and neutralizing antibodies[12–17]. Previous cryo-electron microscopy (cryo-EM) structures of the SARS-CoV-2 S glycoprotein in two distinct functional states[6,9]—along with cryo-EM and crystal structures of the SARS-CoV-2 S^B in complex with ACE2[18–20]—have revealed dynamic states of the S^B domains, providing a blueprint for the design of vaccines and inhibitors of viral entry.

Passive administration of monoclonal antibodies (mAbs) could have a major effect on controlling the SARS-CoV-2 pandemic by providing immediate protection, complementing the development of prophylactic vaccines. Accelerated development of mAbs in a pandemic setting could be reduced to 5–6 months, compared to the traditional timeline of 10–12 months[21]. The recent finding that ansuvimab (mAb114) is a safe and effective treatment for symptomatic infection with Ebola virus is a notable example of the successful use of mAb therapy during an outbreak of infectious disease[22,23]. Potently neutralizing human mAbs

---

[1]Humabs BioMed SA, Vir Biotechnology, Bellinzona, Switzerland. [2]Department of Biochemistry, University of Washington, Seattle, WA, USA. [3]Institut Pasteur and CNRS UMR 3569, Unité de Virologie Structurale, Paris, France. [4]Vir Biotechnology, San Francisco, CA, USA. [5]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [6]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO, USA. [7]Institute for Research in Biomedicine, Università della Svizzera Italiana, Bellinzona, Switzerland. [8]Department of Molecular Microbiology, Washington University School of Medicine, St Louis, MO, USA. [9]These authors contributed equally: Dora Pinto, Young-Jun Park, Martina Beltramello, Alexandra C. Walls. ✉e-mail: dveesler@uw.edu; dcorti@vir.bio

Table 1 | Characteristics of the antibodies described in this study

mAb	VH (per cent identity)	HCDR3 sequence	VL (per cent identity)	SARS-CoV	SARS-CoV-2	Binding
S110	VH3-30 (96.88)	AKDRFQFARSWYGDYFDY	VK2-30 (96.60)	+	+	RBD and non-RBD
S124	VH2-26 (98.28)	ARINTAAYDYDSTTFDI	VK1-39 (98.57)	+	+	RBD
S109	VH3-23 (93.75)	ARLESATQPLGYYFYGMDV	VL3-25 (97.85)	+	−	RBD
S111	VH3-30 (95.14)	ARDIRHLIVVVSDMDV	VK2-30 (98.30)	+	−	RBD
S127	VH3-30 (96.53)	AKDLFGYCRSTSCESLDD	VK1-9 (98.92)	+	−	RBD
S215	VH3-30 (90.28))	ARETRHYSHGLNWFDP	VK3-15 (98.92)	+	−	RBD
S217	VH3-49 (95.58)	SWIHRIVS	VK1-33 (98.21)	+	−	RBD
S218	VH3-30 (93.40)	ARDVKGHIVVMTSLDY	VK2-30 (97.62)	+	−	RBD
S219	VH1-58(92.01)	AAEMATIQNYYYYYGMDV	VK1-39 (95.34)	+	−	RBD
S222	VH1-2 (91.67)	ARGDVPVGTGWVFDF	VK1-39 (92.47)	+	−	RBD
S223	VH3-30 (95.14)	ATVSVEGYTSGWYLGTLDF	VK3-15 (98.21)	+	−	RBD
S224	VH1-18 (90.97)	ARQSHSTRGGWHFSP	VK1-39 (95.70)	+	−	RBD
S225	VH3-9 (96.18)	AKDISLVFWSVNPPRNGMDV	VK1-39 (98.57)	+	−	RBD
S226	VH3-30 (89.61)	ARDSSWQSTGWPINWFDR	VK3-11 (96.11)	+	−	RBD
S227	VH3-23 (95.14)	ASPLRNYGDLLY	VK1-5 (96.06)	+	−	RBD
S228	VH3-30 (96.53)	ARDLQMRVVVVSNFDY	VK2D-30 (99.32)	+	−	RBD
S230	VH3-30 (90.97)	VTQRDNSRDYFPHYFHDMDV	VK2-30 (97.62)	+	−	RBD
S231	VH3-30 (90.62)	ARDDNLDRHWPLRLGGY	VK2-30 (94.56)	+	−	RBD
S237	VH3-21 (96.53)	ARGFERYYFDS	VL1-44 (96.84)	+	−	RBD
S309	VH1-18 (97.22)	ARDYTRGAWFGESLIGGFDN	VK3-20 (97.52)	+	+	RBD
S315	VH3-7 (97.92)	ARDLWWNDQAHYYGMDV	VL3-25 (97.57)	+	+	RBD
S303	VH3-23 (90.28)	ARERDDIFPMGLNAFDI	VK1-5 (97.49)	+	+	RBD
S304	VH3-13 (97.89)	ARGDSSGYYYYFDY	VK1-39 (93.55)	+	+	RBD
S306	VH1-18 (95.49)	ASDYFDSSGYYHSFDY	VK3-11 (98.92)	+	+	Non-RBD
S310	VH1-69 (92.71)	ATRTYDSSGYRPYYYGLDV	VL2-23 (97.57)	+	+	Non-RBD

VH and VL per cent identity refers to V gene segment identity compared to germline (as per the International Immunogenetics Information System (http://www.imgt.org/)).

from the memory B cells of individuals infected with SARS-CoV[24] or Middle East respiratory syndrome coronavirus (MERS-CoV)[25] have previously been isolated. Passive transfer of these mAbs protected mice challenged with various SARS-CoV isolates and SARS-related coronaviruses[24,26,27], as well as with MERS-CoV[25]. Structural characterization of two of these mAbs in complex with the S glycoprotein of SARS-CoV or MERS-CoV provided molecular-level information on the mechanisms of viral neutralization[14]. In particular, although both mAbs blocked S^B attachment to the host receptor, the S230 mAb (which neutralizes SARS-CoV) acted by functionally mimicking attachment to the receptor and promoting fusogenic conformational rearrangements of the S glycoprotein[14]. Another mechanism of SARS-CoV neutralization has recently been described for mAb CR3022, which bound a cryptic epitope that is only accessible when at least two out of the three S^B domains of a S-glycoprotein trimer were in the open conformation[28,29]. However, none of these mAbs neutralizes SARS-CoV-2. A mAb termed 47D11 that neutralizes SARS-CoV and SARS-CoV-2 was also recently isolated from human-immunoglobulin transgenic mice[30], and several mAbs have been isolated from individuals infected with SARS-CoV-2[31].

Identifying a SARS-CoV-2 cross-neutralizing mAb

A set of human neutralizing mAbs (from an individual infected with SARS-CoV in 2003) that potently inhibit both human and zoonotic SARS-CoV isolates has previously been identified[24,26,32]. To characterize the potential cross-reactivity of these antibodies with SARS-CoV-2, we performed a memory B cell screening using peripheral blood mononuclear cells collected in 2013 from the same patient. Here we describe 19 mAbs from the initial screen (2004 blood draw)[24,26] and 6 mAbs from the new screen (2013 blood draw). The mAbs that we identified had a broad usage of V gene segments, and were not clonally related (Table 1). Eight out of the twenty-five mAbs bound to CHO cells that express SARS-CoV-2 S glycoprotein or SARS-CoV S glycoprotein, with half-maximal effective concentration values that ranged between 1.4 and 6,100 ng ml^{-1}, and 0.8 and 254 ng ml^{-1}, respectively (Fig. 1a, b). We further evaluated the mAbs for binding to the SARS-CoV-2 and SARS-CoV S^B domains, as well as to the prefusion-stabilized ectodomain trimers of human coronavirus HCoV-OC43[33], MERS-CoV[34,35], SARS-CoV[35] and SARS-CoV-2[6] S glycoproteins. None of the mAbs that we studied bound to prefusion ectodomain trimers of the HCoV-OC43 or MERS-CoV S glycoproteins, which indicated a lack of cross-reactivity outside the *Sarbecovirus* subgenus (Extended Data Fig. 1). The mAbs S303, S304, S309 and S315 bound SARS-CoV-2 and SARS-CoV RBDs with nano- to sub-picomolar affinity (Extended Data Fig. 2). In particular, the S309 IgG bound to the immobilized SARS-CoV-2 S^B domain and to the ectodomain trimer of the S glycoprotein with sub-picomolar and picomolar avidities, respectively (Fig. 1c). The S309 Fab bound with nanomolar to sub-nanomolar affinities to both molecules (Fig. 1d). S306 and S310 stained cells that express SARS-CoV-2 S glycoprotein at higher levels than cells that express SARS-CoV S glycoprotein, yet these mAbs did not interact with ectodomain trimers and RBD constructs of SARS-CoV-2 or SARS-CoV S glycoprotein by enzyme-linked immunosorbent assay. These results suggest that S306 and S310 may recognize post-fusion SARS-CoV-2 S glycoprotein, which has recently been proposed to be abundant on the surface of authentic SARS-CoV-2 viruses[36] (Fig. 1a, b, Extended Data Fig. 3).

To evaluate the neutralization potency of the SARS-CoV-2 cross-reactive mAbs, we carried out pseudovirus neutralization assays using a murine leukaemia virus (MLV) pseudotyping system[37]. S309 showed comparable neutralization potencies against both

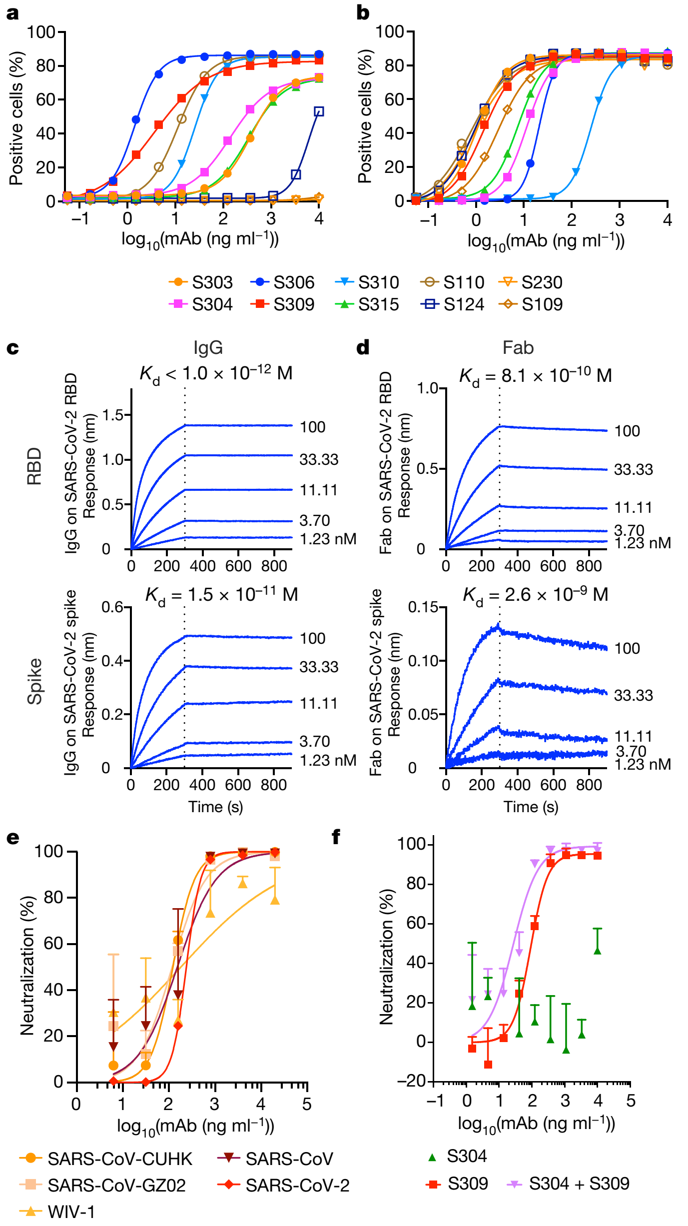


Fig. 1 | Identification of a potent SARS-CoV-2 neutralizing mAb from an individual infected with SARS-CoV. a, **b**, Binding of a panel of mAbs, isolated from a patient immune to SARS-CoV, to SARS-CoV-2 (**a**) or SARS-CoV (**b**) S glycoproteins expressed at the surface of expiCHO cells (symbols are means of duplicates from one experiment). **c**, **d**, Avidity and affinity measurement of S309 IgG1 (**c**) and Fab (**d**) for binding to immobilized SARS-CoV-2 S^B domain (RBD) and to the prefusion ectodomain trimer of S glycoprotein, measured using biolayer interferometry. **e**, Neutralization of SARS-CoV-2–MLV, SARS-CoV–MLV (bearing S glycoprotein from various isolates) and the SARS-related coronavirus WIV-1 by mAb S309. Mean ± s.d. of triplicates is shown for all pseudoviruses, except for SARS-CoV-2 (mean of duplicates). **f**, Neutralization of authentic SARS-CoV-2 (strain n-CoV/USA_WA1/2020) by mAbs as measured by a focus-forming assay on Vero E6 cells. For the cocktail of S309 and S304, the concentration of S309 is as indicated in the *x* axis. S304 was added at a constant amount of 20 μg ml^{-1}. Mean ± s.d. of quadruplicates is shown. In **a**, **b**, all mAbs in the same experiment were tested once. Individual mAbs were tested independently with similar results. In **c**–**f**, one representative out of two experiments with similar results is shown.

SARS-CoV and SARS-CoV-2 pseudoviruses, whereas S303 neutralized SARS-CoV–MLV but not SARS-CoV-2–MLV. S304 and S315 weakly neutralized SARS-CoV–MLV and SARS-CoV-2–MLV (Extended Data Fig. 4).

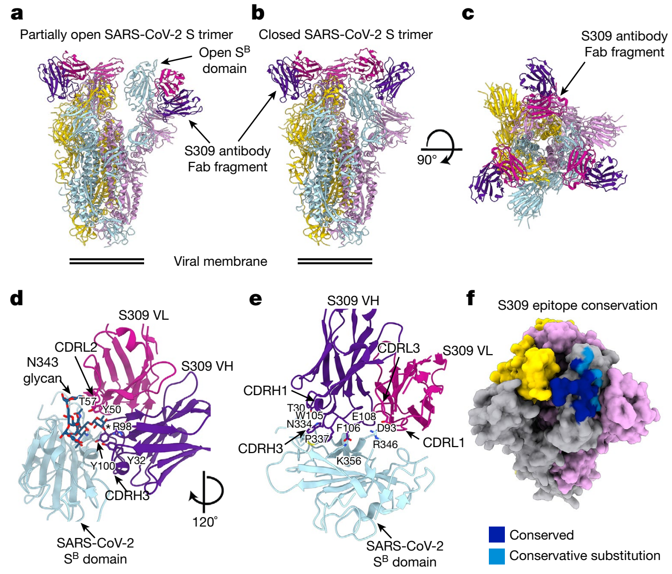


Fig. 2 | Cryo-EM structures of the SARS-CoV-2 S glycoprotein in complex with the S309 neutralizing mAb Fab fragment. a, Ribbon diagram of the partially open SARS-CoV-2 S-glycoprotein trimer (one S^B domain is open) bound to three S309 Fabs. The Fab bound to the open B domain is included only for visualization, and was omitted from the final model. **b**, **c**, Ribbon diagrams in two orthogonal orientations of the closed SARS-CoV-2 S-glycoprotein trimer bound to three S309 Fabs. **d**, Close-up view of the S309 epitope, showing the contacts formed with the core fucose (labelled with a star) and the rest of the glycan at position N343. **e**, Close-up view of the S309 epitope, showing the 20-residue-long CDRH3 sitting atop the S^B helix that comprises residues 337–344. The oligosaccharide at position N343 is omitted for clarity. In **a**–**e**, only the Fab variable domains are shown. In **d**, **e**, selected residues involved in interactions between S309 and SARS-CoV-2 S glycoprotein are shown. **f**, Molecular surface representation of the SARS-CoV-2 S-glycoprotein trimer, showing the S309 epitope on one protomer coloured by residue conservation among SARS-CoV-2 and SARS-CoV S glycoproteins. The other two protomers are coloured pink and gold.

In addition, S309 neutralized SARS-CoV–MLVs from isolates of the 3 phases of the 2002–2003 epidemic with half-maximal inhibitory concentration (IC_{50}) values of between 120 and 180 ng ml^{-1}, and partially neutralized the SARS-related coronavirus[38] WIV-1 (Fig. 1e). Finally, mAb S309 potently neutralized authentic SARS-CoV-2 (2019n-CoV/USA_WA1/2020) with an IC_{50} of 79 ng ml^{-1} (Fig. 1f).

Structural basis of S309 cross-neutralization

To study the mechanisms of S309-mediated neutralization, we characterized the complex between the S309 Fab fragment and a prefusion stabilized ectodomain trimer of SARS-CoV-2 S glycoprotein[6] using single-particle cryo-EM. Similar to a previous study of apo SARS-CoV-2 S glycoprotein[6], 3D classification of the cryo-EM data enabled identification of two structural states: a trimer with one S^B domain open, and a closed trimer. We determined 3D reconstructions at 3.7 Å and 3.1 Å resolution, respectively, of the ectodomain trimer of the SARS-CoV-2 S glycoprotein with a single open S^B domain and in a closed state (applying three-fold symmetry), both with three S309 Fabs bound (Fig. 2a–c, Extended Data Fig. 5a–f). In parallel, we also determined a crystal structure of the S309 Fab at 3.3 Å resolution to assist model building (Extended Data Fig. 5g). The S309 Fab bound to the open S^B domain is weakly resolved in the cryo-EM map, owing to marked conformational variability of the upward pointing S^B domain, and was not modelled in density. The analysis below is based on the closed-state structure.

S309 recognizes a proteoglycan epitope on the SARS-CoV-2 S^B, distinct from the receptor-binding motif. The epitope is accessible in both

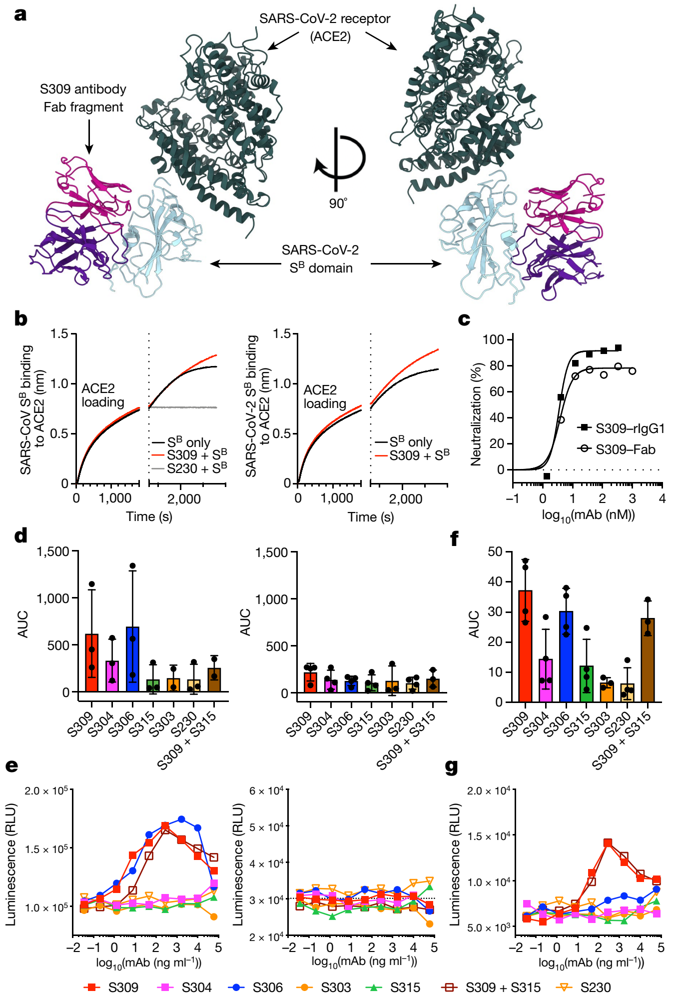


Fig. 3 | Mechanism of S309 neutralization. a, Ribbon diagrams of S309 and ACE2 bound to SARS-CoV-2 S^B. This composite model was generated using the SARS-CoV-2 S glycoprotein–S309 cryo-EM structure reported here, and a previously published crystal structure of SARS-CoV-2 S^B bound to ACE2[20]. Only the Fab variable domains are shown. **b**, Competition of S309 or S230 mAbs with ACE2 to bind to SARS-CoV S^B (left) and SARS-CoV-2 S^B (right). The vertical dashed line indicates the start of the association of mAb-complexed or free S^B to solid-phase ACE2. Each mAb was tested in at least two experiments with similar results. **c**, Neutralization of SARS-CoV–MLV by S309 recombinant (r) IgG1 or S309 Fab, plotted in nM. Mean of duplicates is shown; the experiment was repeated with similar results. **d**, mAb-mediated ADCC using primary natural killer effector cells and SARS-CoV-2 S-glycoprotein-expressing expiCHO as target cells. Bars show the average area under the curve (AUC) for the responses of 3 (VV) (left) or 4 (FF-FV) (right) donors genotyped for their FcγRIII. Mean ± s.d., data were pooled from two independent experiments. **e**, Activation of high-affinity (V158) (left) or low-affinity (F158) (right) FcγRIIIa was measured using Jurkat reporter cells and SARS-CoV-2 S-glycoprotein-expressing expiCHO as target cells. RLU, relative luminescence unit. **f**, mAb-mediated ADCP using Cell-Trace-Violet-labelled peripheral blood mononuclear cells as phagocytic cells, and PKF67-labelled SARS-CoV-2 S-glycoprotein-expressing expiCHO as target cells. Bars show the AUC for the responses of four donors. Mean ± s.d., data were pooled from two independent experiments. **g**, Activation of FcγRIIa (131H) measured using Jurkat reporter cells and SARS-CoV-2 S-glycoprotein-expressing expiCHO as target cells. In **e** and **g**, one experiment, symbols show means of duplicates per mAb dilution except for S304, S230 and S315 (one data point per dilution). S309 was retested in an independent experiment with similar results.

the open and closed states of the S glycoprotein, which explains the stoichiometric binding of Fab to the trimer of the S glycoprotein (Fig. 2a–c). The S309 paratope is composed of all 6 complementarity-determining region (CDR) loops, which bury a surface area of about 1,150 Å^2 at the interface with S^B through electrostatic interactions and hydrophobic contacts. The 20-residue-long CDRH3 sits atop the S^B helix that comprises residues 337–344, and also contacts the edge of the S^B 5-stranded β-sheet (residues 356–361), overall accounting for about 50% of the buried surface area (Fig. 2d, e). CDRL1 and CDRL2 extend the epitope by interacting with the helix that spans residues 440–444, which is located near the three-fold molecular axis of the S glycoprotein. CDRH3 and CDRL2 sandwich the glycan of the SARS-CoV-2 S glycoprotein at position N343, through contacts with the core fucose moiety (consistent with a previous study that detected SARS-CoV-2 N343 core-fucosylated peptides by mass spectrometry[16]) and to a lesser extent with the other saccharides within the glycan chain (Fig. 2d). These interactions between S309 and the glycan bury an average surface of about 300 Å^2 and stabilize the N343 oligosaccharide, which is resolved to a much larger extent than in structures of the apo SARS-CoV-2 S glycoprotein[6,9].

The structural data explain the S309 cross-reactivity between SARS-CoV-2 and SARS-CoV, as 17 out of 22 residues of the epitope are strictly conserved (Fig. 2f, Extended Data Fig. 6a, b). R346, N354, R357 and L441 of SARS-CoV-2 are conservatively substituted for K333, E341, K344 (except for SARS-CoV isolate GZ02, in which this is R444) and I428, respectively, of SARS-CoV, and the K444 of SARS-CoV-2 is semi-conservatively substituted for T431 of SARS-CoV, in agreement with the comparable binding affinities of S309 to SARS-CoV and SARS-CoV-2 S glycoprotein (Fig. 1c). The oligosaccharide at position N343 is also conserved in both viruses and corresponds to SARS-CoV N330, for which core-fucosylated glycopeptides were previously detected by mass spectrometry[14] that would allow for similar interactions with the S309 Fab. Analysis of the S glycoprotein sequences of the 11,839 SARS-CoV-2 isolates reported to date indicates that the epitope residues are conserved in all but 4 isolates, for which we found N354D or the S359N substitutions that are not expected to affect recognition by S309 (Extended Data Fig. 7a, b). Furthermore, S309 contact residues are highly conserved across human and animal isolates of clade 1, 2 and 3 sarbecoviruses[39] (Extended Data Fig. 7c). Collectively, our data suggest that S309 could neutralize potentially all SARS-CoV-2 isolates known to be circulating to date, and possibly many other zoonotic sarbecoviruses. The degree of conservation is consistent with the moderate rates of evolution of SARS-CoV-2, estimated at about 1.8×10^{-3} substitutions per site per year[40]. On the basis of more than 10^4 viral sequences analysed to date, an estimated 112 residues are under positive selection (8 in the S glycoprotein) and 18 are under negative selection (1 in the S glycoprotein) in a genome of nearly 30 kb[41]. These observations are consistent with the fact that *Coronaviridae* is a family of RNA viruses with unusually high replication fidelity required by their exceptionally large genomes[42].

Mechanism of S309-mediated neutralization

The cryo-EM structure of S309 bound to SARS-CoV-2 S glycoprotein presented here, combined with the structures of SARS-CoV-2 S^B and SARS-CoV S^B in complex with ACE2, indicate[18–20,43] that the Fab engages an epitope distinct from the receptor-binding motif and would not clash with ACE2 upon binding to S glycoprotein (Fig. 3a). Biolayer interferometry analysis of S309 Fab or IgG binding to the SARS-CoV-2 S^B domain or the ectodomain trimer of S glycoprotein confirmed the absence of competition between S309 and ACE2 for binding to the SARS-CoV-2 S glycoprotein (Fig. 3b, Extended Data Fig. 8).

To further investigate the mechanism of S309-mediated neutralization, we compared side-by-side infection of SARS-CoV-2–MLV in the presence of either S309 Fab or S309 IgG. Both experiments yielded comparable IC_{50} values (3.8 and 3.5 nM, respectively), indicating similar potencies for IgG and Fab (Fig. 3c). However, S309 IgG-mediated neutralization reached 100%, whereas neutralization plateaued at about 80% in the presence of S309 Fab (Fig. 3c). This result indicates that one

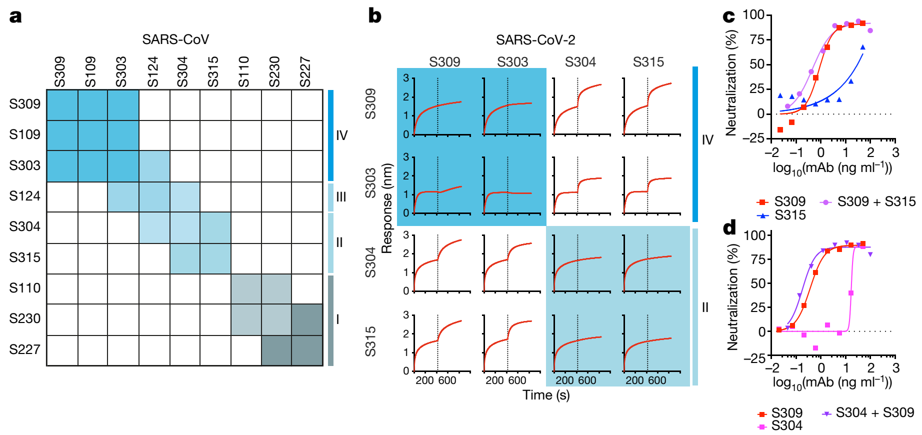


Fig. 4 | mAb cocktails enhance SARS-CoV-2 neutralization. **a**, Heat map showing the competition of mAb pairs for binding to the SARS-CoV S^B domain as measured by biolayer interferometry (as shown in Extended Data Fig. 10). **b**, Competition of mAb pairs for binding to the SARS-CoV-2 S^B domain. Each competition measured once. In **a**, **b**, a coloured box indicates competition and each colour corresponds to an antigenic site. **c**, **d**, Neutralization of SARS-CoV-2–MLV by S309 combined with an equimolar amount of S304 or S315 mAb. For mAb cocktails, the concentration on the *x* axis is that of the total amount of mAb per dilution. Mean of duplicates is shown. Both experiments were repeated with similar results.

or more IgG-specific bivalent mechanisms—such as S-glycoprotein trimer cross-linking, steric hindrance or aggregation of virions[44]—may contribute to the ability of S309 to fully neutralize pseudovirions.

Fc-dependent effector mechanisms, such as antibody-dependent cell cytotoxicity (ADCC) mediated by natural killer cells, can contribute to viral control in individuals infected with virus. We observed efficient S309- and S306-mediated ADCC of SARS-CoV-2 S-glycoprotein-transfected cells, whereas the other mAbs that we tested showed limited or no activity (Fig. 3d, Extended Data Fig. 9a). These findings might be related to distinct binding orientations and/or positioning of the mAb Fc fragment relative to the FcγRIIIa receptors. ADCC was observed only using natural killer (effector) cells that express the high-affinity FcγRIIIa variant (V158) but not the low-affinity variant (F158) (Fig. 3d). These results, which we confirmed using a FcγRIIIa cell reporter assay (Fig. 3e), suggest that S309 Fc engineering could potentially enhance the activation of natural killer cells with the low-affinity FcγRIIIa variant (F158)[45]. Antibody-dependent cellular phagocytosis (ADCP), mediated by macrophages or dendritic cells, can contribute to viral control by clearing virus and infected cells and by stimulating a T cell response via presentation of viral antigens[46,47]. Similar to the ADCC results, the mAbs S309 and S306 showed the strongest ADCP response (Fig. 3f, Extended Data Fig. 9b). However, FcγRIIa signalling was observed only for S309 (Fig. 3g). These findings suggest that ADCP by monocytes was dependent on engagement of both FcγRIIIa and FcγRIIa. Collectively, these results demonstrate that—in addition to potent in vitro neutralization—S309 may leverage additional protective mechanisms in vivo, as has previously been shown for other antiviral antibodies[48,49]. Although the risks of antibody-dependent disease enhancement will need to be evaluated for SARS-CoV-2, potent virus neutralization by a specific monoclonal antibody or by antibody cocktails is expected to limit this possibility, compared to weakly neutralizing antibodies that might potentially be induced upon vaccination or infection[50]. Furthermore, neutralizing antibodies are associated with reduced susceptibility to re-infection or disease in humans[51–53]. Regardless, the possibility of antibody-dependent disease enhancement will need to be assessed during clinical trials for either antibodies or vaccines.

Enhancing SARS-CoV-2 neutralization

To gain more insight into the epitopes recognized by our panel of mAbs, we used structural information, escape mutant analysis[26,32,35] and biolayer-interferometry-based epitope binning to map the antigenic sites that are present on the SARS-CoV and SARS-CoV-2 S^B domains (Fig. 4a, Extended Data Fig. 10). This analysis identified at least four antigenic sites within the S^B domain of SARS-CoV targeted by our panel of mAbs. The receptor-binding motif, which is targeted by S230, S227 and S110, is termed site I. Sites II and III are defined by S315 and S124, respectively, and the two sites were bridged by mAb S304. Site IV is defined by the S309, S109 and S303 mAbs. Given the lower number of mAbs that cross-react with SARS-CoV-2, we were able to identify site IV targeted by S309 and S303, and sites II and III targeted by S304 and S315 (Fig. 4b).

On the basis of these findings, we evaluated the neutralization potency of the site-IV S309 mAb in combination with either the site-II S315 mAb or the site-II and site-III S304 mAb. Although S304 and S315 alone were weakly neutralizing, the combination of either of these mAbs with S309 resulted in an enhanced neutralization potency, compared to single mAbs, against both SARS-CoV-2–MLV and authentic SARS-CoV-2 (Figs. 1f, 4c, d). A synergistic effect between two non-competing anti-RBD mAbs has previously been reported for SARS-CoV[54] and our data extend this observation to SARS-CoV-2, providing a proof-of-concept for the use of mAb combinations to prevent or control SARS-CoV-2.

In summary, our study identifies S309 as a human mAb that has broad neutralizing activity against multiple sarbecoviruses (including SARS-CoV-2), via recognition of a highly conserved epitope in the S^B domain that comprises the N343 glycan (N330 in SARS-CoV S glycoprotein). Furthermore, S309 can recruit effector mechanisms and showed increased neutralization in combination with weakly neutralizing mAbs, which may mitigate the risk of viral escape. Our data indicate the potential to discover potently neutralizing pan-sarbecovirus mAbs, define antigenic sites to include in vaccine design and pave the way to support preparedness for future outbreaks of sarbecoviruses. As S309 shows promise as an effective countermeasure to the COVID-19 pandemic caused by SARS-CoV-2, Fc variants of S309 with increased half-life and effector functions have entered an accelerated development path towards clinical trials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2349-y.

1. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
2. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med*. **382**, 727–733 (2020).
3. Tortorici, M. A. & Veesler, D. Structural insights into coronavirus entry. *Adv. Virus Res*. **105**, 93–116 (2019).
4. Walls, A. C. et al. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531**, 114–117 (2016).
5. Walls, A. C. et al. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc. Natl Acad. Sci. USA* **114**, 11157–11162 (2017).
6. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
7. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
8. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol*. **5**, 562–569 (2020).
9. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
10. Li, W. et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J*. **24**, 1634–1643 (2005).
11. Rossen, J. W. et al. The viral spike protein is not involved in the polarized sorting of coronaviruses in epithelial cells. *J. Virol*. **72**, 497–503 (1998).
12. Walls, A. C. et al. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nat. Struct. Mol. Biol*. **23**, 899–905 (2016).
13. Xiong, X. et al. Glycan shield and fusion activation of a deltacoronavirus spike glycoprotein fine-tuned for enteric infections. *J. Virol*. **92**, e01628-17 (2018).
14. Walls, A. C. et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell* **176**, 1026–1039 (2019).
15. Yang, Y. et al. Two mutations were critical for bat-to-human transmission of Middle East respiratory syndrome coronavirus. *J. Virol*. **89**, 9119–9123 (2015).
16. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific analysis of the SARS-CoV-2 glycan spike. *Science* https://doi.org/10.1126/science.abb9983 (2020).
17. Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl Acad. Sci. USA* **114**, E7348–E7357 (2017).
18. Yan, R. et al. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
19. Shang, J. et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
20. Lan, J. et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
21. Kelley, B. Developing therapeutic monoclonal antibodies at pandemic pace. *Nat. Biotechnol*. **38**, 540–545 (2020).
22. Corti, D. et al. Protective monotherapy against lethal Ebola virus infection by a potently neutralizing antibody. *Science* **351**, 1339–1342 (2016).
23. Levine, M. M. Monoclonal antibody therapy for Ebola virus disease. *N. Engl. J. Med*. **381**, 2365–2366 (2019).
24. Traggiai, E. et al. An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med*. **10**, 871–875 (2004).
25. Corti, D. et al. Prophylactic and postexposure efficacy of a potent human monoclonal antibody against MERS coronavirus. *Proc. Natl Acad. Sci. USA* **112**, 10473–10478 (2015).
26. Rockx, B. et al. Structural basis for potent cross-neutralizing human monoclonal antibody protection against lethal human and zoonotic severe acute respiratory syndrome coronavirus challenge. *J. Virol*. **82**, 3220–3235 (2008).
27. Menachery, V. D. et al. SARS-like WIV1-CoV poised for human emergence. *Proc. Natl Acad. Sci. USA* **113**, 3048–3053 (2016).
28. Yuan, M. et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630–633 (2020).
29. Joyce, M. G. et al. A cryptic site of vulnerability on the receptor binding domain of the SARS-CoV-2 spike glycoprotein. Preprint at bioRxiv https://doi.org/10.1101/2020.03.15.992883 (2020).
30. Wang, C. et al. A human monoclonal antibody blocking SARS-CoV-2 infection. *Nat. Commun*. **11**, 2251 (2020).
31. Ju, B., Zhang, Q., Ge, J. et al. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* https://doi.org/10.1038/s41586-020-2380-z (2020).
32. Rockx, B. et al. Escape from human monoclonal antibody neutralization affects in vitro and in vivo fitness of severe acute respiratory syndrome coronavirus. *J. Infect. Dis*. **201**, 946–955 (2010).
33. Tortorici, M. A. et al. Structural basis for human coronavirus attachment to sialic acid receptors. *Nat. Struct. Mol. Biol*. **26**, 481–489 (2019).
34. Park, Y. J. et al. Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. *Nat. Struct. Mol. Biol*. **26**, 1151–1157 (2019).
35. Walls, A. C. et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell* **176**, 1026–1039.e15 (2019).
36. Liu, C. et al. Viral architecture of SARS-CoV-2 with post-fusion spike revealed by cryo-EM. Preprint bioRxiv https://doi.org/10.1101/2020.03.02.972927 (2020).
37. Millet, J. K. & Whittaker, G. R. Murine leukemia virus (MLV)-based coronavirus spike-pseudotyped particle production and infection. *Bio. Protoc*. **6**, e2035 (2016).
38. Ge, X. Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
39. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
40. Li, X. et al. Transmission dynamics and evolutionary history of 2019-nCoV. *J. Med. Virol*. **92**, 501–511 (2020).
41. Pond, S. SARS-CoV-2 Natural Selection Analysis. http://data.hyphy.org/web/covid-19/dashboard/ (2020).
42. Smith, E. C., Sexton, N. R. & Denison, M. R. Thinking outside the triangle: replication fidelity of the largest RNA viruses. *Annu. Rev. Virol*. **1**, 111–132 (2014).
43. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868 (2005).
44. Klasse, P. J. & Sattentau, Q. J. Occupancy and mechanism in antibody-mediated neutralization of animal viruses. *J. Gen. Virol*. **83**, 2091–2108 (2002).
45. Wang, X., Mathieu, M. & Brezski, R. J. IgG Fc engineering to modulate antibody effector functions. *Protein Cell* **9**, 63–73 (2018).
46. He, W. et al. Alveolar macrophages are critical for broadly-reactive antibody-mediated protection against influenza A virus in mice. *Nat. Commun*. **8**, 846 (2017).
47. DiLillo, D. J. & Ravetch, J. V. Differential Fc–receptor engagement drives an anti-tumor vaccinal effect. *Cell* **161**, 1035–1045 (2015).
48. Corti, D. et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850–856 (2011).
49. Hessell, A. J. et al. Fc receptor but not complement binding is important in antibody protection against HIV. *Nature* **449**, 101–104 (2007).
50. Iwasaki, A. & Yang, Y. The potential danger of suboptimal antibody responses in COVID-19. *Nat. Rev. Immunol*. **20**, 339–341 (2020).
51. Callow, K. A. Effect of specific humoral immunity and some non-specific factors on resistance of volunteers to respiratory coronavirus infection. *J. Hyg. (Lond.)* **95**, 173–189 (1985).
52. Reed, S. E. The behaviour of recent isolates of human respiratory coronavirus in vitro and in volunteers: evidence of heterogeneity among 229E-related strains. *J. Med. Virol*. **13**, 179–192 (1984).
53. Callow, K. A., Parry, H. F., Sergeant, M. & Tyrrell, D. A. The time course of the immune response to experimental coronavirus infection of man. *Epidemiol. Infect*. **105**, 435–446 (1990).
54. ter Meulen, J. et al. Human monoclonal antibody combination against SARS coronavirus: synergy and coverage of escape mutants. *PLoS Med*. **3**, e237 (2006).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines

Cell lines used in this study were obtained from ATCC (HEK293T, Vero-E6, Vero CCL81 cells) or Invitrogen (Expi-CHO cells). All cell lines used in this study were routinely tested for mycoplasma and found to be mycoplasma-free.

Ethics statement

Donors provided written informed consent for the use of blood and blood components (such as sera), following approval by the Canton Ticino Ethics Committee (Switzerland).

Antibody discovery and expression

mAbs were isolated from Epstein–Barr-virus-immortalized memory B cells. Recombinant antibodies were expressed in expiCHO cells transiently cotransfected with plasmids expressing the heavy and light chain, as previously described[55]. The mAbs S303, S304, S306, S309, S310 and S315 were expressed as recombinant IgG-LS antibodies. The LS mutation confers a longer half-life in vivo[56]. Antibodies S110 and S124 tested in Fig. 1 and Extended Data Fig. 1 were purified mAbs produced from immortalized B cells.

Transient expression of recombinant SARS-CoV-2 protein and flow cytometry

The full-length *S* gene of SARS-CoV-2 strain (SARS-CoV-2-S) isolate BetaCoV/Wuhan-Hu-1/2019 (accession number MN908947) was codon-optimized for human cell expression and cloned into the phCMV1 expression vector (Genlantis). Expi-CHO cells were transiently transfected with phCMV1-SARS-CoV-2-S, SARS-spike_pcDNA-3.1 (strain SARS) or empty phCMV1 (mock) using Expifectamine CHO Enhancer. Two days after transfection, cells were collected for immunostaining with mAbs. An Alexa-647-labelled secondary antibody anti-human IgG Fc was used for detection. Binding of mAbs to transfected cells was analysed by flow cytometry using a ZE5 Cell Analyzer (Biorard) and FlowJo software (TreeStar). Positive binding was defined by differential staining of CoV-S-glycoprotein transfectants versus mock transfectants.

Affinity and avidity determination and competition experiments using Octet (biolayer interferometry)

For affinity and avidity determination of IgG1 compared to Fab fragment, biotinylated RBD of SARS-CoV-2 (produced in house; residues 331–550 of S glycoprotein from BetaCoV/Wuhan-Hu-1/2019, accession number MN908947, biotinylated with EZ-Link NHS–PEG$_4$–biotin from ThermoFisher) and biotinylated SARS-CoV-2 2P S glycoprotein avi-tagged were loaded at 7.5 μg/ml in kinetics buffer (0.01% endotoxin-free BSA, 0.002% Tween-20, 0.005% NaN3 in PBS) for 8 min onto streptavidin biosensors (Molecular Devices, ForteBio). Association of IgG1 and Fab was performed in kinetics buffer at 100, 33, 11, 3.6, 1.2 nM for 5 min. Dissociation in kinetics buffer was measured for 10 min. K_d values were calculated using a 1:1 global fit model (Octet).

Alternatively, measurement of apparent K_d for IgGs was determined using protein A biosensors (Pall ForteBio) that were loaded with different mAbs at 2.7 μg/ml for 1 min, after a hydration step for 10 min in kinetics buffer. Association curves were recorded for 5 min by incubating the mAb-coated sensors with different concentrations of SARS-CoV RBD (Sino Biological) or SARS-CoV-2 RBD (produced in house; residues 331–550 of S glycoprotein from BetaCoV/Wuhan-Hu-1/2019, accession number MN908947). The highest RBD concentration was 10 μg/ml, then serially diluted 1:2.5. Dissociation was recorded for 9 min by moving the sensors to wells containing kinetics buffer. K_d values were calculated using a global fit model (Octet). Octet Red96 (ForteBio) equipment was used.

For mAb competition experiments, His-tagged RBD of SARS-CoV or SARS-CoV-2 was loaded for 5 min at 3 μg/ml in kinetics buffer onto anti-Penta-HIS (HIS1K) biosensors (Molecular Devices, ForteBio). Association of mAbs was performed in kinetics buffer at 15 μg/ml.

For ACE2 competition experiments, ACE2–His (Bio-Techne AG) was loaded for 30 min at 5 μg/ml in kinetics buffer onto anti-HIS (HIS2) biosensors (Molecular Devices-ForteBio).

SARS-CoV RBD–rabbit Fc or SARS-CoV-2 RBD–mouse Fc (Sino Biological Europe GmbH) at 1 μg/ml was associated for 15 min, after preincubation with or without antibody (30 μg/ml, 30 min). Dissociation was monitored for 5 min.

Enzyme-linked immunosorbent assay

The following proteins were coated on 96-well enzyme-linked immunosorbent assay plates at the following concentrations: SARS-CoV RBD (Sino Biological, 40150-V08B1) at 1 μg/ml, SARS-CoV-2 RBD (produced in house) at 10 μg/ml, ectodomains (stabilized prefusion trimer) of SARS-CoV, SARS-CoV-2, HCoV-OC43 and MERS-CoV, all at 1 μg/ml. After blocking with 1% BSA in PBS, antibodies were added to the plates at concentrations between 5 and 0.000028 μg/ml and incubated for 1 h at room temperature. Plates were washed and secondary antibody Goat Anti Human IgG-AP (Southern Biotechnology: 2040-04) was added. Substrate *p*-nitrophenyl pPhosphate (pNPP) (Sigma-Aldrich 71768) was used for colour development. Optical density at 405 nm was read on an ELx808IU plate reader (Biotek).

Measurement of Fc-mediated effector functions

ADCC assays were performed using expiCHO cells transiently transfected with SARS-CoV or SARS-CoV-2 S glycoprotein as targets. Natural killer cells were isolated from fresh blood of healthy donors using the MACSxpress NK Isolation Kit (Miltenyi Biotec, cat. no. 130-098-185). Target cells were incubated with titrated concentrations of mAbs for 10 min and were then incubated with primary human natural killer cells as effector cells at an effector:target ratio of 9:1. ADCC was measured using LDH release assay (Cytotoxicity Detection Kit (LDH) (Roche; cat. no. 11644793001)) after 4 h incubation at 37 °C.

ADCP assays were performed using expiCHO target cells transiently transfected with SARS-CoV-2 S glycoprotein and fluorescently labelled with PKH67 Fluorescent Cell Linker Kits (Sigma Aldrich, cat. no. MINI67). Target cells were incubated with titrated concentrations of mAbs for 10 min, followed by incubation with human peripheral blood mononuclear cells isolated from healthy donors and fluorescently labelled with Cell Trace Violet (Invitrogen, cat. no. C34557). An effector:target ratio of 20:1 was used. After an overnight incubation at 37 °C, cells were stained with anti-human CD14-APC antibody (BD Pharmingen, cat. no. 561708, Clone M5E2) to stain monocytes. Antibody-mediated phagocytosis was determined by flow cytometry, gating on CD14$^+$ cells that were double-positive for cell trace violet and PKH67.

Determination of mAb-dependent activation of human FcγRIIIa or FcγRIIa was performed using expiCHO cells transiently transfected with SARS-CoV-2 S glycoprotein (BetaCoV/Wuhan-Hu-1/2019). Cells were incubated with titrated concentrations of mAbs for 10 min before incubation with FcγRIIIa-receptor- or FcγRIIa-expressing Jurkat cells stably transfected with NFAT-driven luciferase gene (Promega, cat. no. G9798 and G7018). An effector-to-target ratio of 6:1 for FcγRIIIa and 5:1 for FcγRIIa was used. Activation of human FcγR in this bioassay results in the NFAT-mediated expression of a luciferase reporter gene. Luminescence was measured after 21 h of incubation at 37 °C with 5% CO$_2$, using the Bio-Glo-TM Luciferase Assay Reagent according to the manufacturer's instructions.

Pseudovirus neutralization assays

MLV-based SARS-CoV S-glycoprotein-pseudotyped viruses were prepared as previously described[6,37]. HEK293T cells were cotransfected with a SARS-CoV, SARS-CoV-2, CUHK, GZ02 or WiV1 S-glycoprotien-encoding-plasmid, an MLV Gag-Pol packaging construct and the MLV transfer vector encoding a luciferase reporter using the Lipofectamine 2000 transfection reagent (Life Technologies) according to the manufacturer's instructions. Cells were incubated for 5 h at 37 °C with 8% CO_2 with OPTIMEM transfection medium. DMEM containing 10% FBS was added for 72 h.

VeroE6 cells or DBT cells transfected with human ACE2 were cultured in DMEM containing 10% FBS, 1% penicillin–streptomycin and plated into 96-well plates for 16–24 h. Concentrated pseudovirus with or without serial dilution of antibodies was incubated for 1 h and then added to the wells after washing 3× with DMEM. After 2–3 h DMEM containing 20% FBS and 2% penicillin–streptomycin was added to the cells for 48 h. Following 48 h of infection, One-Glo-EX (Promega) was added to the cells and incubated in the dark for 5–10 min before reading on a Varioskan LUX plate reader (ThermoFisher). Measurements were done in duplicate and relative luciferase units were converted to per cent neutralization and plotted with a nonlinear regression curve fit in PRISM.

Live virus neutralization assay

SARS-CoV-2 strain 2019-nCoV/USA_WA1/2020 was obtained from the Centers for Disease Control and Prevention (gift of N. Thornburg). Virus was passaged once in Vero CCL81 cells (ATCC) and titrated by focus-forming assay on Vero E6 cells. Serial dilutions of the indicated mAbs were incubated with 10^2 focus-forming units of SARS-CoV-2 for 1 h at 37 °C. Mab–virus complexes were added to Vero E6 cell monolayers in 96-well plates and incubated at 37 °C for 1 h. Subsequently, cells were overlaid with 1% (w/v) methylcellulose in MEM supplemented with 2% FBS. Plates were collected 30 h later by removing overlays and fixed with 4% PFA in PBS for 20 min at room temperature. Plates were washed and sequentially incubated with 1 µg/ml of CR3022[57] anti-S-glycoprotein antibody and HRP-conjugated goat anti-human IgG in PBS supplemented with 0.1% saponin and 0.1% BSA. SARS-CoV-2-infected cell foci were visualized using TrueBlue peroxidase substrate (KPL) and quantified on an ImmunoSpot microanalyzer (Cellular Technologies). Data were processed using Prism software (GraphPad Prism 8.0).

Recombinant S-glycoprotein ectodomain and S^B production

The SARS-CoV-2 2P S glycoprotein (GenBank: YP_009724390.1) ectodomain was produced in 500-ml cultures of HEK293F cells grown in suspension using FreeStyle 293 expression medium (Life technologies) at 37 °C in a humidified 8% CO_2 incubator rotating at 130 r.p.m., as previously reported[6]. The culture was transfected using 293fectin (ThermoFisher Scientific) with cells grown to a density of 10^6 cells per ml and cultivated for 3 d. The supernatant was collected and cells were resuspended for another three days, yielding two collections. Clarified supernatants were purified using a 5-ml Cobalt affinity column (Takara). Purified protein was concentrated and flash-frozen in a buffer containing 20 mM Tris pH 8.0 and 150 mM NaCl before cryo-EM analysis. The SARS-CoV-2 2P S-glycoprotein-avi, SARS-CoV S glycoprotein, HCoV-OC43 S glycoprotein and MERS-CoV S glycoprotein constructs have previously been described[14,33] and were produced similarly to SARS-CoV-2 2P S glycoprotein.

Cryo-EM sample preparation and data collection

Three microlitres of SARS-CoV-2 S glycoprotein at 1.6 mg/ml was mixed with 0.45 µl of S309 Fab (obtained by LysC fragmentation of S309 IgG) at 7.4 mg/ml for 1 min at room temperature before application onto a freshly glow-discharged 1.2/1.3 UltraFoil grid (300 mesh). Plunge freezing used a vitrobot MarkIV (ThermoFisher Scientific) using a blot force of 0 and 6.5 s blot time at 100% humidity and 25 °C. Data were acquired using the Leginon software[58] to control an FEI Titan Krios transmission electron microscope operated at 300 kV and equipped with a Gatan K2 Summit direct detector and Gatan Quantum GIF energy filter, operated in zero-loss mode with a slit width of 20 eV. Automated data collection was carried out using Leginon at a nominal magnification of 130,000× with a pixel size of 0.525 Å with tilt angles ranging between 20° and 50°, as previously described[59]. The dose rate was adjusted to 8 counts per pixel per s, and each movie was acquired in super-resolution mode fractionated in 50 frames of 200 ms. Three thousand nine hundred micrographs were collected in a single session with a defocus range of between −1.0 and −3.0 µm.

Cryo-EM data processing

Movie frame alignment, estimation of the microscope contrast-transfer function parameters, particle picking and extraction were carried out using Warp[60]. Particle images were extracted with a box size of 800 binned to 400, yielding a pixel size of 1.05 Å. For each dataset, two rounds of reference-free 2D classification were performed using cryoSPARC[61] to select well-defined particle images. Subsequently, two rounds of 3D classification with 50 iterations each (angular sampling 7.5° for 25 iterations and 1.8° with local search for 25 iterations), using a previously reported closed SARS-CoV-2 S glycoprotein structure[6] as initial model, were carried out using Relion[62] without imposing symmetry to separate distinct SARS-CoV-2 S glycoprotein conformations. Three-dimensional refinements were carried out using non-uniform refinement along with per-particle defocus refinement in cryoSPARC[61]. Particle images were subjected to Bayesian polishing[63] before performing another round of non-uniform refinement in cryoSPARC[61], followed by per-particle defocus refinement and again non-uniform refinement. Reported resolutions are based on the gold-standard Fourier shell correlation of 0.143 criterion and Fourier shell correlation curves were corrected for the effects of soft masking by high-resolution noise substitution[64].

Cryo-EM model building and analysis

UCSF Chimera[65] and Coot were used to fit atomic models (Protein Data Bank codes (PDB) 6VXX and PDB 6VYB) into the cryo-EM maps. The Fab was subsequently manually built using Coot[66,67]. N-linked glycans were hand-built into the density where visible, and the models were refined and relaxed using Rosetta[68]. Glycan refinement relied on a dedicated Rosetta protocol, which uses physically realistic geometries based on prior knowledge of saccharide chemical properties[69], and was aided by using both sharpened and unsharpened maps. Models were analysed using MolProbity[70], EMringer[71], Phenix[72] and privateer[73] to validate the stereochemistry of both the protein and glycan components. Figures were generated using UCSF ChimeraX[74].

Crystallization and X-ray structure determination of Fab S309

Fab S309 crystals were grown in a hanging drop set up with a mosquito at 20 °C using 150 nl protein solution in Tris HCl pH 8.0, 150 mM NaCl and 150 nl mother liquor solution containing 1.1 M sodium malonate, 0.1 M HEPES, pH 7.0 and 0.5% (w/v) Jeffamine ED-2001. Crystals were cryo-protected using the mother liquor solution supplemented with 30% glycerol. The dataset was collected at ALS beamline 5.0.2 and processed to 3.3 Å resolution in space group $P4_12_12$ using mosflm[75] and Aimless[76]. The structure of Fab S309 was solved by molecular replacement using Phaser[77] and homology models as search models. The coordinates were improved and completed using Coot[66] and refined with REFMAC5[78]. Crystallographic data collection and refinement statistics are shown in Extended Data Fig. 5g.

Conservation analysis

SARS-CoV-2 genomics sequences were downloaded from GISAID on 27 April 2020 (n=11,839) using the 'complete (>29,000 bp)' and 'low coverage exclusion' filters. Bat and pangolin sequences were removed to yield human sequences only. The S glycoprotein ORF was localized by aligning the reference protein sequence (YP_009724390.1) to the genomic sequence of isolates with Exonerate v.2.4.0 (-m protein2dna --refine full --minintron 999999 --percent 30 --showalignment false --showvulgar false --ryo “>%ti\n%tcs). Coding nucleotide sequences were translated in silico using seqkit v.0.12.0. Multiple sequence alignment was performed using MAFFT v.7.455 (--amino–bl 80 --nomemsave --reorder --add spike_aa_sequences.fasta --keeplength reference_aa_sequence.fasta). Variants were determined by comparison of aligned sequences to the reference sequence using the R v3.6.3/Bioconductor v.3.10 package Biostrings v.2.54.0 (function: consensusMatrix). A similar strategy was used to extract and translate S glycoprotein sequences from SARS-CoV genomes sourced from ViPR (search criteria: SARS-related coronavirus, full-length genomes, human host, deposited before December 2019 to exclude SARS-CoV-2, $n = 53$, performed on 29 March 2020). We confirmed that sourced SARS-CoV genome sequences comprised all the major published strains (such as Urbani, Tor2, TW1, P2, and Frankfurt1, among others). Pangolin sequences[79] were sourced from GISAID and bat sequences from the three clades of sarbecoviruses[39] were downloaded from Genbank (civet (AY304486.1) and raccoon dog (AY304487.1). Full conservation analysis code is available at https://github.com/virbio/manuscript-cov2-pinto-conservation.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-EM maps and atomic models have been deposited at the Electron Microscopy Data Bank and the PDB with accession codes EMD-21864 and 6WPS (closed SARS-CoV-2 S in complex with S309), as well as EMD-21865 and 6WPT (SARS-CoV-2 S with one S^B open in complex with S309). The crystal structure of the S309 Fab was deposited to the PDB with accession code 6WS6. Materials generated in this study will be made available on request, but we may require a completed materials transfer agreement.

55. Stettler, K. et al. Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. *Science* **353**, 823–826 (2016).
56. Zalevsky, J. et al. Enhanced antibody half-life improves in vivo activity. *Nat. Biotechnol*. **28**, 157–159 (2010).
57. Tian, X. et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg. Microbes Infect*. **9**, 382–385 (2020).
58. Suloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol*. **151**, 41–60 (2005).
59. Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
60. Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat. Methods* **16**, 1146–1152 (2019).
61. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
62. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
63. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5–17 (2019).
64. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).
65. Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J. Struct. Biol*. **157**, 281–287 (2007).
66. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
67. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
68. Wang, R. Y. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016).
69. Frenz, B. et al. Automatically fixing errors in glycoprotein structures with Rosetta. *Structure* **27**, 134–139.e3 (2019).
70. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
71. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
72. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
73. Agirre, J. et al. Privateer: software for the conformational validation of carbohydrate structures. *Nat. Struct. Mol. Biol*. **22**, 833–834 (2015).
74. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci*. **27**, 14–25 (2018).
75. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).
76. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69**, 1204–1214 (2013).
77. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr*. **40**, 658–674 (2007).
78. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
79. Lam, T. T. et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* https://doi.org/10.1038/s41586-020-2169-0 (2020).

Acknowledgements This study was supported by the National Institute of General Medical Sciences (R01GM120553, D.V.), the National Institute of Allergy and Infectious Diseases (HHSN272201700059C to D.V. and 75N93019C00062 to M.S.D.)), a Pew Biomedical Scholars Award (D.V.), an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund (D.V.), the University of Washington Arnold and Mabel Beckman cryo-EM centre, the Pasteur Institute (M.A.T.) and beamline 5.0.2 at the Advanced Light Source at Lawrence Berkley National Laboratory. Beamline 5.0.2 is supported by The Berkeley Center for Structural Biology. The Advanced Light Source is a Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231. We thank N. Zheng for use of his crystallization robot. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database upon which this research is based.

Author contributions A.C.W., K.F., M.S.D., D.V. and D.C. designed the experiments. A.C.W., M.A.T., S.J. and E.C. expressed and purified the proteins. K.C., F.Z., S.J. and E.C. sequenced and expressed antibodies. D.P., M.B., A.C.W. and S.B. performed binding assays. D.P., M.B., A.C.W., A.P. and A.D.M. carried out pseudovirus neutralization assays. J.B.C. and R.E.C. performed neutralization assays with authentic SARS-CoV-2. B.G. performed effector function assays. Y.-J.P. prepared samples for cryo-EM and collected the data. Y.-J.P. and D.V. processed the data, built and refined the atomic models. A.C.W. crystallized the S309 Fab. Y.-J.P. collected and processed the X-ray diffraction data and built and refined the atomic model. R.S., A.T. and G.S. performed bioinformatic and conservation analysis. A.L. provided key reagents. A.C.W., K.F., C.H.-D., H.W.V., A.L., D.V. and D.C. analysed the data and prepared the manuscript with input from all authors.

Competing interests D.P., S.B., K.C., E.C., C.H.-D., G.S., M.B., A.K., K.F., A.P., F.Z., S.J., B.G., A.D.M., A.L., A.T., H.W.V., R.S. and D.C. are employees of Vir Biotechnology and may hold shares in Vir Biotechnology. M.S.D. is a consultant for Inbios, Eli Lilly, Vir Biotechnology, NGM Biopharmaceuticals and Emergent BioSolutions, and on the Scientific Advisory Board of Moderna. The Diamond laboratory at Washington University School of Medicine has received sponsored research agreements from Moderna. The other authors declare no competing interests.

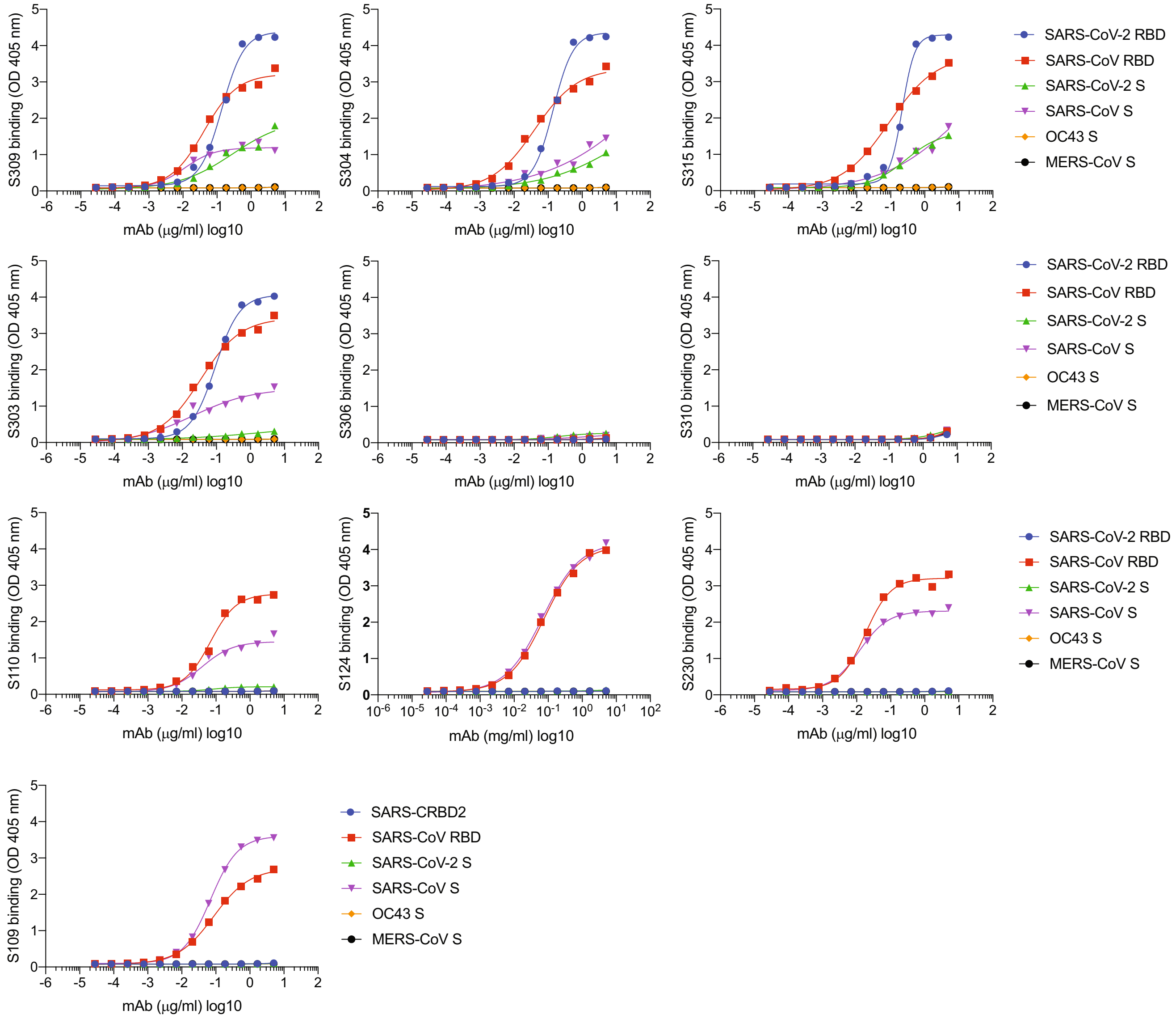
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2349-y.

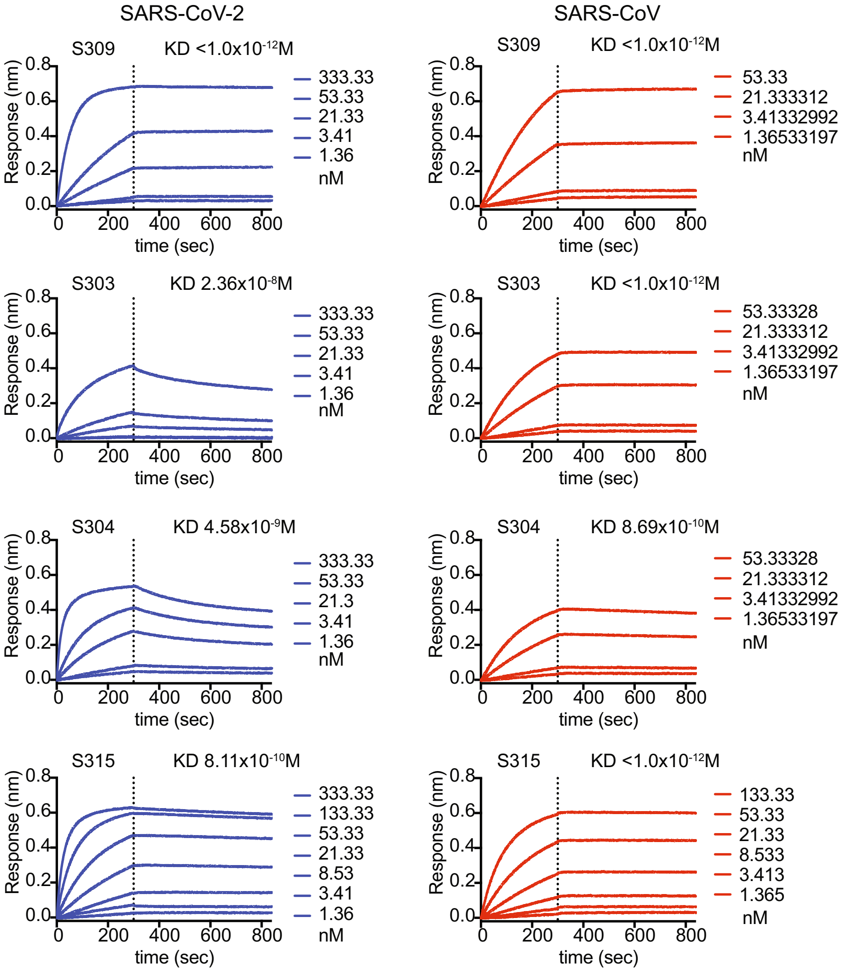
Correspondence and requests for materials should be addressed to D.V. or D.C.

Peer review information *Nature* thanks Pamela Bjorkman, Michel Nussenzweig and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

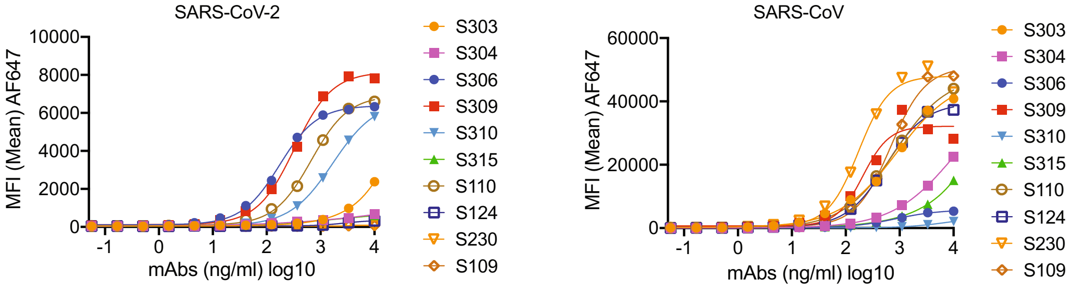
Reprints and permissions information is available at http://www.nature.com/reprints.



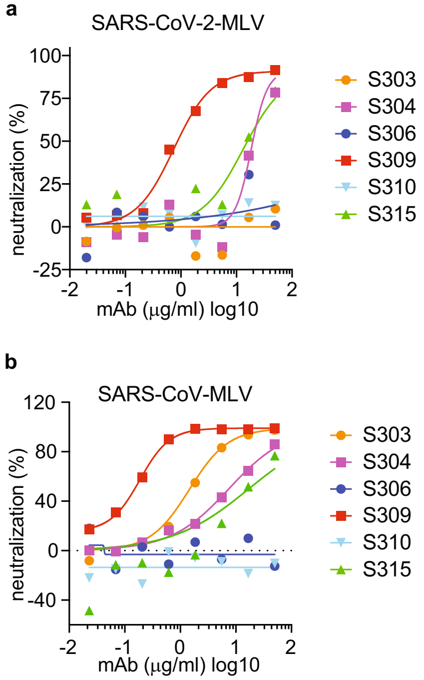
Extended Data Fig. 1 | Binding of cross-reactive antibodies to RBD of SARS-CoV and SARS-CoV-2, and to ectodomains of different coronavirus strains. mAbs were tested by enzyme-linked immunosorbent assay at a concentration range of 5 to 0.00028 μg ml^{-1}. S, stabilized prefusion trimer of the indicated coronavirus. Symbols show mean of duplicates. The full dataset was tested once. Binding of S309 to the RBDs and all S glycoproteins was repeated in an independent experiment with similar results.



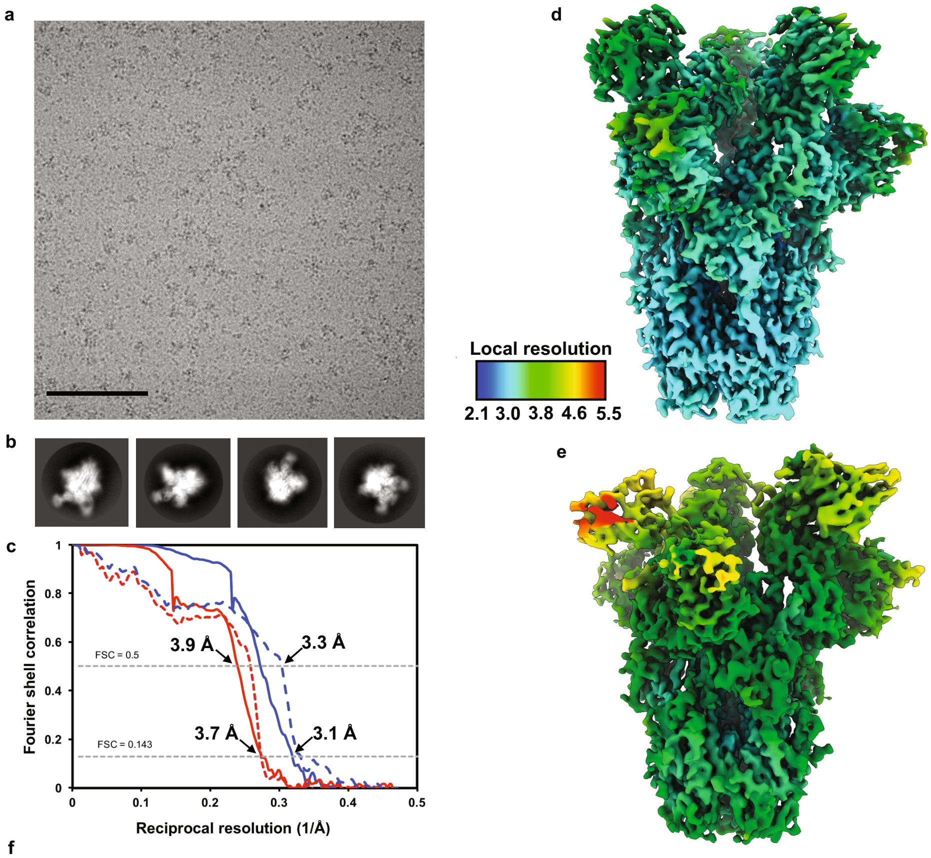
Extended Data Fig. 2 | Antibody affinity and avidity of S309, S303, S304 and S315 to the RBD of SARS-CoV and SARS-CoV-2. Antibodies were loaded to biolayer interferometry (BLI) pins via protein A for the measurement of association of different concentrations of the RBD of SARS-CoV-2 (blue) and SARS-CoV (red). Vertical dashed lines indicate the start of the dissociation phase, when BLI pins were switched to buffer. The experiments were done once for S303, S304 and S315. The experiment for S309 was repeated once.



Extended Data Fig. 3 | Binding of crossreactive mAbs to expiCHO cells transfected with SARS-CoV or SARS-CoV-2 S glycoprotein. Mean fluorescence intensity as measured in flow cytometry for each antibody. Antibody concentrations tested are indicated in the *x* axis.



Extended Data Fig. 4 | Neutralization of cross-reactive mAbs. a, **b**, The six cross-reactive mAbs indicated in the legend were tested for neutralization of SARS-CoV-2–MLV (**a**) and SARS-CoV–MLV (**b**). Symbols are mean of duplicates. **a**, One experiment. Neutralization of S304, S315 and S309 was tested in at least two independent experiments with similar results. **b**, One out of two experiments with similar results.



f

	SARS-CoV-2 S + S309 (closed) (EMD-21864) (PDB ID 6WPS)	SARS-CoV-2 S + S309 (open) (EMD-21865) (PDB ID 6WPT)
Data collection and processing		
Magnification	130,000	130,000
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	70	70
Defocus range (μm)	0.5-3.0	0.5-3.0
Pixel size (Å)	0.525	0.525
Symmetry imposed	C3	C1
Initial particle images (no.)	637,730	637,730
Final particle images (no.)	168,449	119,608
Map resolution (Å)	3.1	3.7
FSC threshold	0.143	0.143
Map resolution range (Å)	2.5-6	3-8
Refinement		
Initial model used (PDB code)	6VXX	6VYB
Model resolution (Å)	3.3	3.9
FSC threshold	0.5	0.5
Map sharpening B factor (Å^2)	-75	-51
Model composition		
Non-hydrogen atoms	28,116	24,326
Protein residues	3,540	3,277
Ligands (glycans)	75	65
B factors (Å^2)		
Protein	38.04	74.84
Ligand (glycans)	36.97	76.97
R.m.s. deviations		
Bond lengths (Å)	0.012	0.011
Bond angles (°)	1.209	1.204
Validation		
MolProbity score	0.68	0.66
Clashscore	0.24	0.43
Poor rotamers (%)	0.1	0.09
Ramachandran plot		
Favored (%)	97.56	97.96
Allowed (%)	2.44	1.95
Disallowed (%)	0.00	0.09

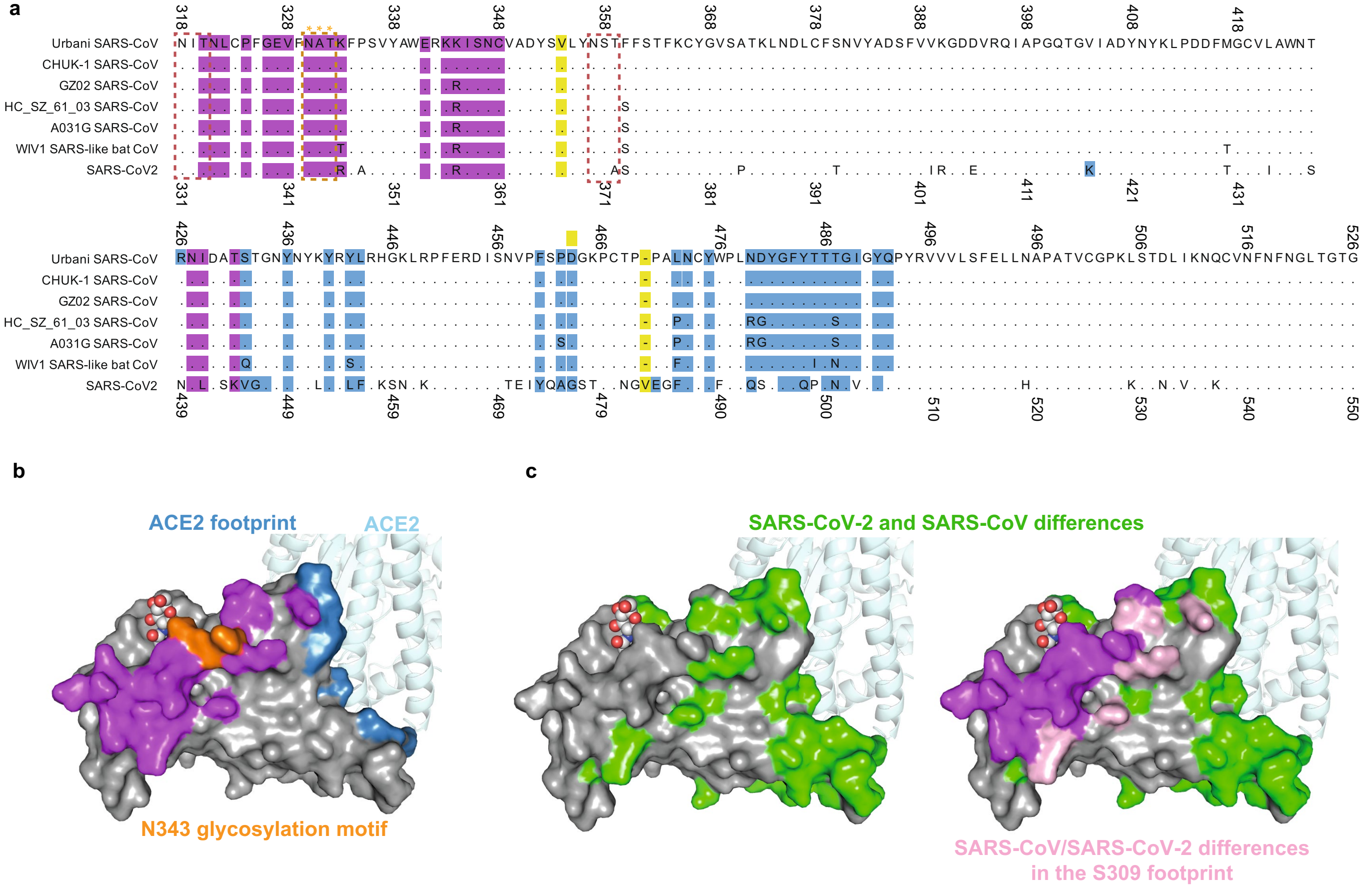
g

	S309 Fab (6WS6)
Data collection	
Space group	P41212
Cell dimensions	
a, *b*, *c* (Å)	132.6, 132.6, 301.2
α, β, γ (°)	90, 90, 90
Resolution (Å)	93.76-3.3 (3.48-3.3) *
R_{sym} or R_{merge}	0.18 (0.75)
$I / \sigma I$	13.2 (3.5)
Completeness (%)	99.4 (99)
Redundancy	12 (11)
Refinement	
Resolution (Å)	93.76-3.3
No. reflections	41,291
R_{work} / R_{free}	0.188/0.227
No. atoms	
Protein	9,630
Ligand	41
B-factors	
Protein	77.9
Ligand	84.2
R.m.s. deviations	
Bond lengths (Å)	0.002
Bond angles (°)	0.42

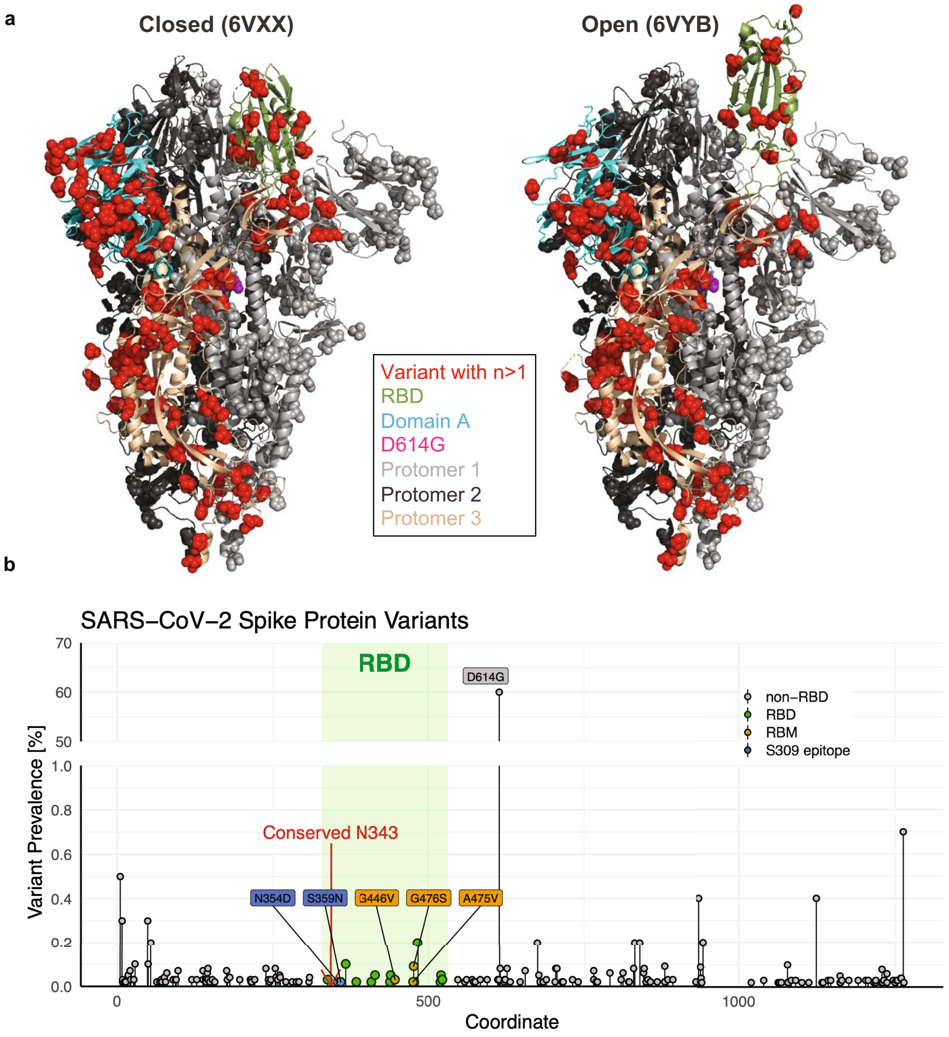
*Number of xtals for each structure should be noted in footnote

*Values in parentheses are for highest-resolution shell

Extended Data Fig. 5 | Cryo-EM data processing and validation. **a**, **b**, Representative electron micrograph (**a**) and class averages (**b**) of SARS-CoV-2 S glycoprotein embedded in vitreous ice. Scale bar, 100 nm. **c**, Gold-standard (solid line) and map/model (dashed line) Fourier shell correlation curves for the closed (blue) and partially open trimers (red). The 0.143 and 0.5 cutoffs are indicated by horizontal grey dashed lines. **d**, **e**, Local-resolution maps calculated using cryoSPARC for the closed- (**d**) and open- (**e**) state reconstructions. **f**, Cryo-EM data collection, refinement and validation statistics. **g**, Data collection and refinement statistics for the S309 Fab X-ray structure.



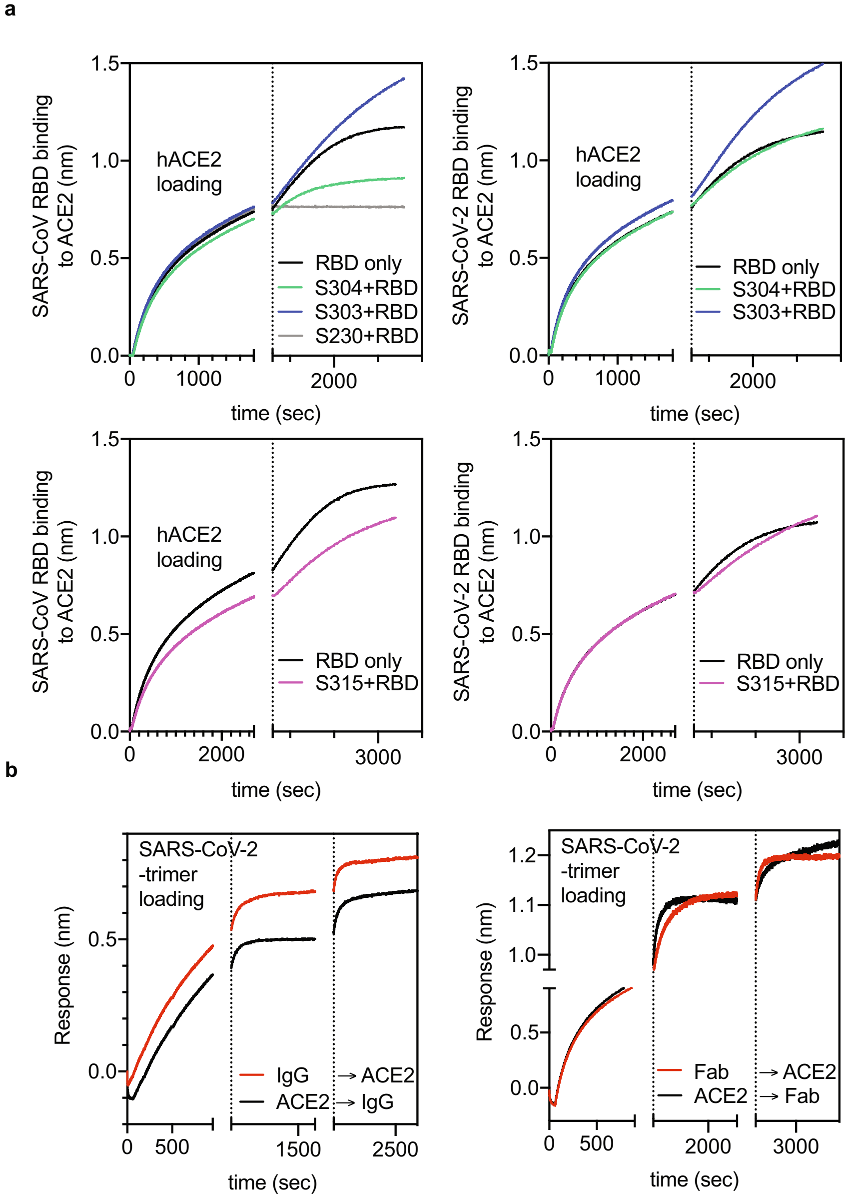
Extended Data Fig. 6 | Conservation of spike and RBD residues. **a**, Alignment of the full RBD of multiple sarbecoviruses. The ACE2 and S309 footprints are highlighted in blue and magenta, respectively. Dashed boxes indicate glycosylation sites and orange asterisks indicate the N343 glycosylation site. **b**, ACE2 and S309 footprints mapped onto the structure of the SARS-CoV-2 RBD domain (PDB 6M0J). The ACE2 footprint was defined by residues being within 5 Å of the receptor in 6M0J. The highly conserved NAT glycosylation motif is shown in orange. **c**, The SARS-CoV and SARS-CoV-2 differences are shown in green, and differences within the S309 footprint are shown in pink (right). The high-frequency RBD variants are shown in yellow (second from left; 4 out of 8 are visible). Red spheres indicate the position of the N343 glycan.



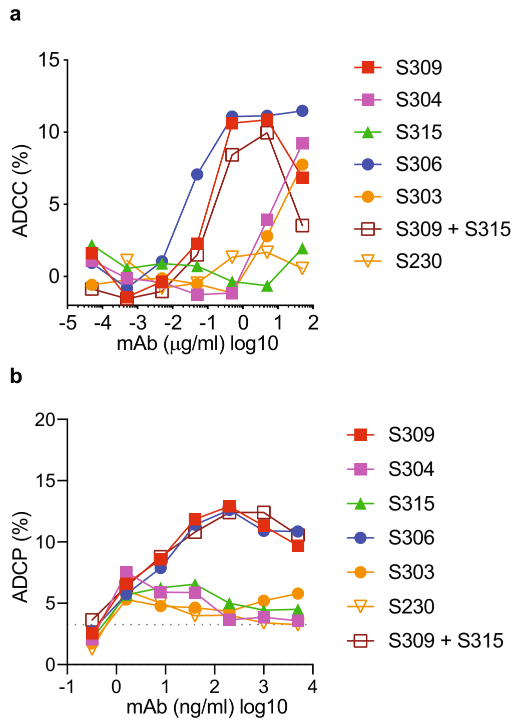
c

		SARS-CoV-2 pos.	333	334	335	337	339	340	341	343	344	345	346	354	356	357	358	359	360	361	440	441	444	509	n	IC50 (μg/ml)
		SARS-CoV pos.	320	321	322	324	326	327	328	330	331	332	333	341	343	344	345	346	347	348	427	428	431	495		
Sarbecoviruses	Clade 2	**Human SARS-CoV-2**	**T**	**N**	**L**	**P**	**G**	**E**	**V**	**N**	**A**	**T**	**R**	**N** (D)	**K**	**R**	**I**	**S** (N)	**N**	**C**	**N**	**L**	**K**	**R**	11839	0.24
		GD Pangolin	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	2	
		Bat CoV RaTG13	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	H	I	-	-		
		GX Pangolin	-	-	-	-	-	-	-	-	-	S	K	-	-	-	-	-	-	-	K	Q	L	-	6	
		Bat CoVs ZC45, ZXC21	-	-	V	-	H	K	-	-	-	-	-	E	T	K	-	-	D	-	K	Q		-	2	
	Clade 3	Human SARS-CoV	-	-	-	-	-	-	-	-	-	-	K	E	-	K/R	-	-	-	-	-	I	T	-	53	0.12-0.15
		Civet SARS-CoV SZ3	-	-	-	-	-	-	-	-	-	-	K	E	-	-	-	-	-	-	-	I	T	-		
		Raccoon Dog SARS-CoV SZ13	-	-	-	-	-	-	-	-	-	-	K	E	-	-	-	-	-	-	-	I	T	-		
		Bat CoV WIV16	-	-	-	-	-	-	-	-	-	-	-	E	-	-	-	-	-	-	-	I	T	-		
		Bat CoV WIV1	-	-	-	-	-	-	-	-	-	-	T	E	-	-	-	-	-	-	-	I	T	-		0.18
		Other Bat CoVs	-	-	R/-	-	D/-	K/-/R	-/I	-	-/V	-/S	-/T	E	T/-/I	K/-	-	-	D/-/E	-	K/-/Q	Q/I/H	G/T/S	-	21	
	Clade 1	Bat CoV BtKY72	-	-	-	-	-	Q	-	-	-	S	N	E	L	-	-	-	D	-	S	V	-	-		
		Bat CoV BGR2008	-	Q	-	-	N	-	-	-	I	-	S	E	M	-	-	T	-	-	S	-		-		

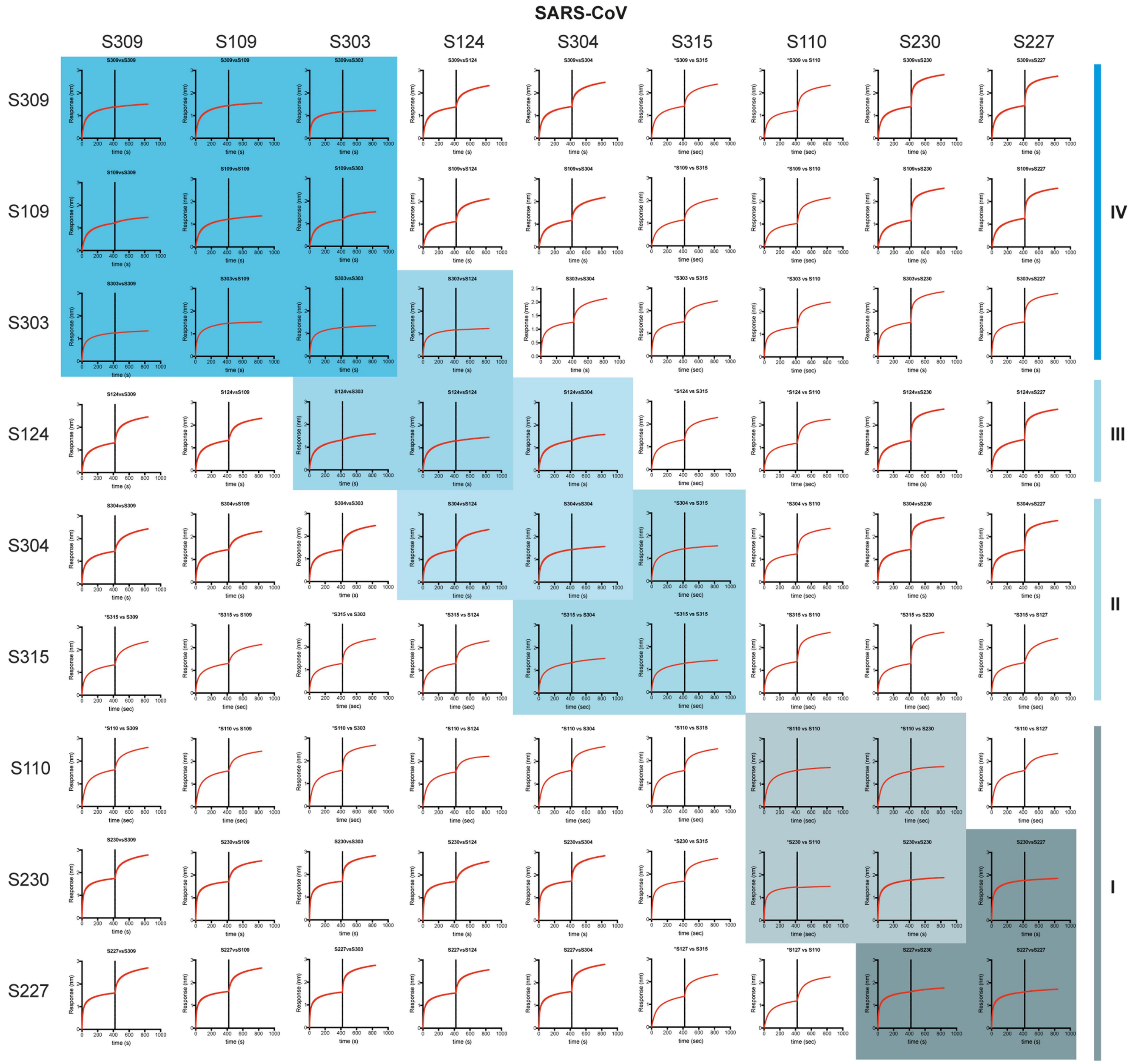
Extended Data Fig. 7 | Conservation of S glycoprotein and RBD residues. SARS-CoV-2 genome sequences retrieved from GISAID on 27 April 27 2020 (*n* = 11,839) were used to annotate variants of the S glycoprotein. Only variants supported by at least 2 sequences (prevalence greater than 0.01%) are shown. **a**, Residues with variants are rendered as red spheres mapped onto the closed and open form of the full trimeric ectodomain (*n* = 171). **b**, Variants are plotted by amino acid position. Each dot is a distinct variant. Colour encodes location (RBD, RBM or S309 epitope). Variants are labelled if their prevalence is greater than 1% (only D614G) or if they belong to the RBM or the S309 epitope. The location of conserved N343 is indicated. **c**, Conservation of residues making contact with S309 across clades of *Sarbecovirus*. Summary of residue variation observed in human and animal SARS-related coronaviruses, arranged by clades[39]. Residue coordinates for both SARS-CoV-2 and SARS-CoV are shown. The NAT motif that directs glycosylation of N343 is highlighted in orange. Substitutions between SARS-CoV and SARS-CoV-2 residues are marked in pink. Dashes indicate identity to SARS-CoV-2 consensus residues. Blanks indicate deletions. When multiple variants are found, they are listed in order of prevalence (from high to low). When *n* > 10,000 (SARS-CoV-2), variants found in a single sequence are not shown, and variants found in 2 sequences only are parenthesized. For the 21 sequences of clade-3 bat CoV-2 from ref. [39] at most the top-3 variants at each position are shown. *n*, number of analysed sequences. IC_{50}, half maximal neutralizing concentration of S309 against SARS-CoV-2–MLV or SARS-CoV.



Extended Data Fig. 8 | Competition of antibodies with RBD binding to ACE2. **a**, Human ACE2 (hACE2) was loaded onto BLI sensors, followed by incubation of the sensors with RBD alone or RBD in combination with recombinant antibodies. The vertical dashed line indicates the start of the loading of RBD with or without antibody. **b**, SARS-CoV-2 ectodomain was loaded onto BLI sensors, followed by incubation of the sensors with macaque ACE2 or S309 IgG (left) or Fab (right). In a third step, sensors were incubated with macaque ACE2 or S309 IgG or Fab as indicated. The experiments were conducted once.



Extended Data Fig. 9 | ADCC and ADCP data for one representative donor. This figure is related to Fig. 3. **a**, ADCC for one donor who is homozygous for high-affinity variant FcγRIIIa 158V (VV). Background signal of cells without antibody was deducted from all values before plotting. **b**, ADCP for one donor who is heterozygous for FcγRIIIa 158V (FV). The dashed line indicates the background signal for cells without antibody. Symbols are mean of duplicates. Cells from each donor were sufficient to conduct one experiment.



Extended Data Fig. 10 | Competition of mAb pairs for binding to the RBD domain of SARS-CoV, as determined by BLI (Octet). RBD was loaded on BLI pins. Association was measured first for antibodies indicated on the left of the matrix, followed by association of the antibodies indicated on top of the matrix. Each combination of antibodies was tested once.

Structural cells are key regulators of organ-specific immune responses

https://doi.org/10.1038/s41586-020-2424-4

Received: 26 August 2019

Accepted: 12 May 2020

Thomas Krausgruber[1,5], Nikolaus Fortelny[1,5], Victoria Fife-Gernedl[1], Martin Senekowitsch[1], Linda C. Schuster[1,3], Alexander Lercher[1], Amelie Nemc[1], Christian Schmidl[1,4], André F. Rendeiro[1], Andreas Bergthaler[1] & Christoph Bock[1,2] ✉

The mammalian immune system implements a remarkably effective set of mechanisms for fighting pathogens[1]. Its main components are haematopoietic immune cells, including myeloid cells that control innate immunity, and lymphoid cells that constitute adaptive immunity[2]. However, immune functions are not unique to haematopoietic cells, and many other cell types display basic mechanisms of pathogen defence[3–5]. To advance our understanding of immunology outside the haematopoietic system, here we systematically investigate the regulation of immune genes in the three major types of structural cells: epithelium, endothelium and fibroblasts. We characterize these cell types across twelve organs in mice, using cellular phenotyping, transcriptome sequencing, chromatin accessibility profiling and epigenome mapping. This comprehensive dataset revealed complex immune gene activity and regulation in structural cells. The observed patterns were highly organ-specific and seem to modulate the extensive interactions between structural cells and haematopoietic immune cells. Moreover, we identified an epigenetically encoded immune potential in structural cells under tissue homeostasis, which was triggered in response to systemic viral infection. This study highlights the prevalence and organ-specific complexity of immune gene activity in non-haematopoietic structural cells, and it provides a high-resolution, multi-omics atlas of the epigenetic and transcriptional networks that regulate structural cells in the mouse.

The structure of most tissues and organs in the mammalian body is shaped by epithelial cells (epithelium), which create internal and external surfaces and barriers; endothelial cells (endothelium), which form the lining of blood vessels; and fibroblasts, which provide essential connective tissue (stroma)[6]. These three cell types, which we refer to as structural cells, have been shown to contribute in important ways to mammalian immunity[7–11]. However, there has been little systematic investigation across organs, in part because structural cells are difficult to study by genetic ablation owing to their essential structural roles in most organs. Multi-omics profiling has emerged as a promising approach to dissecting immune regulation in a systematic, genome-wide manner, as illustrated by recent work on the systems immunology of haematopoietic immune cells[12,13].

In this study, we used multi-omics profiling and integrative bioinformatics to establish a high-resolution atlas of structural cells and of non-haematopoietic immune regulation in the mouse. We observed widespread expression of immune regulators and cytokine signalling molecules in structural cells, organ-specific adaptation to the tissue environment, and unexpectedly diverse capabilities for interacting with haematopoietic cells. These cell-type-specific and organ-specific differences in immune gene activity were reflected by characteristic patterns of chromatin regulation. Notably, we found evidence of an epigenetically encoded immune potential under homeostatic conditions, and the affected genes were preferentially upregulated in response to an immunological challenge induced by systemic viral infection. We validated and functionally dissected this epigenetic potential of structural cells by further in vivo experiments with recombinant cytokines.

In summary, our study uncovered widespread immune gene regulation in structural cells of the mouse, and it established a multi-organ atlas of the underlying epigenetic and transcription-regulatory programs.

Mapping structural cells across organs

To investigate the regulation of immune genes in structural cells, we performed multi-omics profiling of endothelium, epithelium and fibroblasts from 12 mouse organs (brain, caecum, heart, kidney, large intestine, liver, lung, lymph node, skin, small intestine, spleen and thymus). Single-cell suspensions were analysed by flow cytometry, and sort-purified cell populations were profiled with three genome-wide assays (Fig. 1a): (i) gene expression profiling by low-input RNA sequencing (RNA-seq)[14]; (ii) chromatin accessibility profiling with the assay for transposase-accessible chromatin using

---

[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. [2]Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria. [3]Present address: Division of Chromatin Networks, German Cancer Research Center (DKFZ) and Bioquant, Heidelberg, Germany. [4]Present address: Regensburg Center for Interventional Immunology (RCI), Regensburg, Germany. [5]These authors contributed equally: Thomas Krausgruber, Nikolaus Fortelny. ✉e-mail: cbock@cemm.oeaw.ac.at

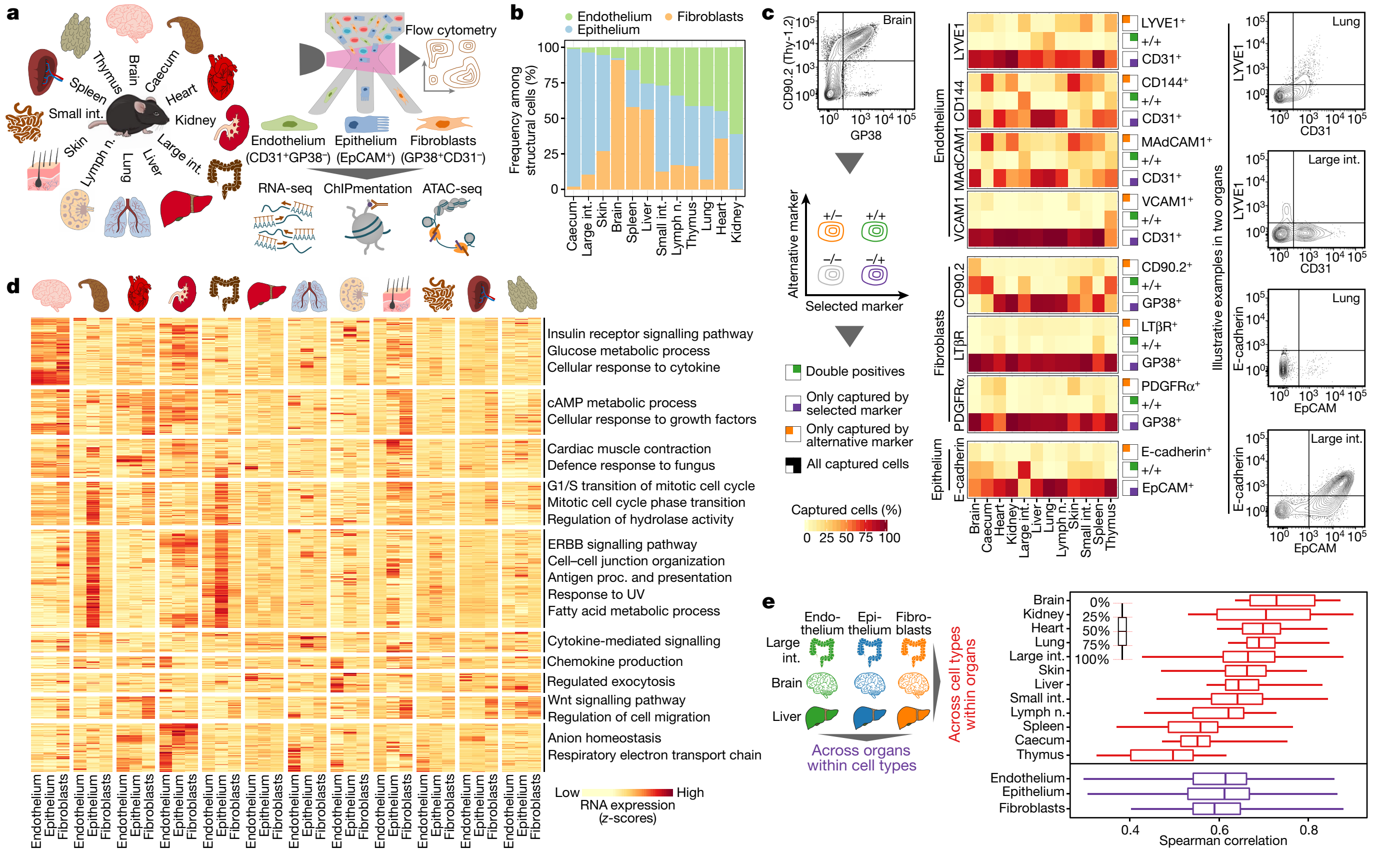


Fig. 1 | Multi-omics profiling establishes cell-type-specific and organ-specific characteristics of structural cells. a, Schematic outline of the experimental approach. **b**, Relative frequencies of structural cell types based on flow cytometry. **c**, Expression of surface markers among structural cells, comparing the standardized sorting of endothelium, epithelium and fibroblasts to potential alternative markers (left: schematic outline; centre: heat maps showing marker overlap; right: illustrative FACS plots). **d**, Expression of differentially regulated genes across cell types and organs. Gene clusters are annotated with enriched terms based on gene set analysis. **e**, Correlation of gene expression across cell types and organs. Sample size: $n = 4$ (**b**) and $n = 3$ (**c**–**e**) independent biological replicates.

sequencing (ATAC-seq)[15]; and (iii) epigenome profiling by ChIPmentation[16] with an antibody against the promoter and enhancer-linked histone H3K4me2 mark[17]. All assays produced high-quality data (Supplementary Table 1). The full dataset is provided as an online resource for interactive browsing and download at http://structural-immunity.computational-epigenetics.org.

To maximize comparability across organs, we developed a standardized workflow for tissue dissociation and cell purification (Fig. 1a). Structural cells were purified with an organ-independent sorting scheme that comprised the endothelium marker CD31 (encoded by *Pecam1*), the epithelium marker EpCAM (*Epcam*) and the fibroblast marker GP38 (encoded by *Pdpn*, also known as podoplanin) (Extended Data Fig. 1a, b). All three types of structural cells were detectable in all 12 organs, with strong differences in their relative frequencies (Fig. 1b, Extended Data Fig. 1c). Our standardized tissue dissociation did not cause major technical biases (Extended Data Fig. 1d).

We validated and phenotypically characterized structural cell populations by flow cytometry for additional markers (Fig. 1c, Extended Data Fig. 2). Purification of CD31^{+}GP38^{-} endothelial cells specifically enriched for blood endothelium while excluding the less prevalent CD31^{+}GP38^{+} cells of lymphatic endothelium, which facilitates the comparison across organs. Alternative endothelial markers such as CD144 (VE-cadherin), MAdCAM1 and VCAM1 (used to assess baseline activation of endothelium under homeostatic conditions) did not improve the identification of blood endothelial cells in our analysis. Similarly, sorting of fibroblasts as GP38^{+}CD31^{-} cells did not miss any major cell populations identified by the alternative markers CD90.2 (Thy-1.2), LTβR or PDGFRα. Finally, E-cadherin could not enhance or replace our sorting of epithelium as EpCAM^{+} cells.

The cellular identity of the sorted endothelium, epithelium and fibroblasts was further confirmed by RNA-seq data analysis, which showed the expected organ-specific (Extended Data Fig. 3a) and cell-type-specific (Extended Data Fig. 3b) patterns of gene expression compared to published multi-tissue expression profiles[18,19]. We also observed the expected expression patterns for various cell-type-specific marker genes, but with a high degree of transcriptional heterogeneity across organs (Extended Data Fig. 3c, Supplementary Table 2). Notably, the expression profiles of structural cells within the same organ were globally more similar to each other than structural cells of the same type across organs (Fig. 1d, e, Extended Data Fig. 3d), which suggests that the tissue and organ environment has a major effect on the transcriptomes of structural cells. Our multi-omics profiles for endothelium, epithelium and fibroblasts across 12 mouse organs thus uncover a marked degree of organ-specific differences among structural cells.

Immune gene activity in structural cells

On the basis of our RNA-seq dataset, we investigated the immune gene activity in structural cells. First, we assessed the ability of structural cells to communicate with haematopoietic immune cells, by inferring a network of potential cell–cell interactions based on known receptor–ligand pairs. The resulting cell–cell interaction network (Fig. 2a, Supplementary Table 3) predicted frequent crosstalk between structural

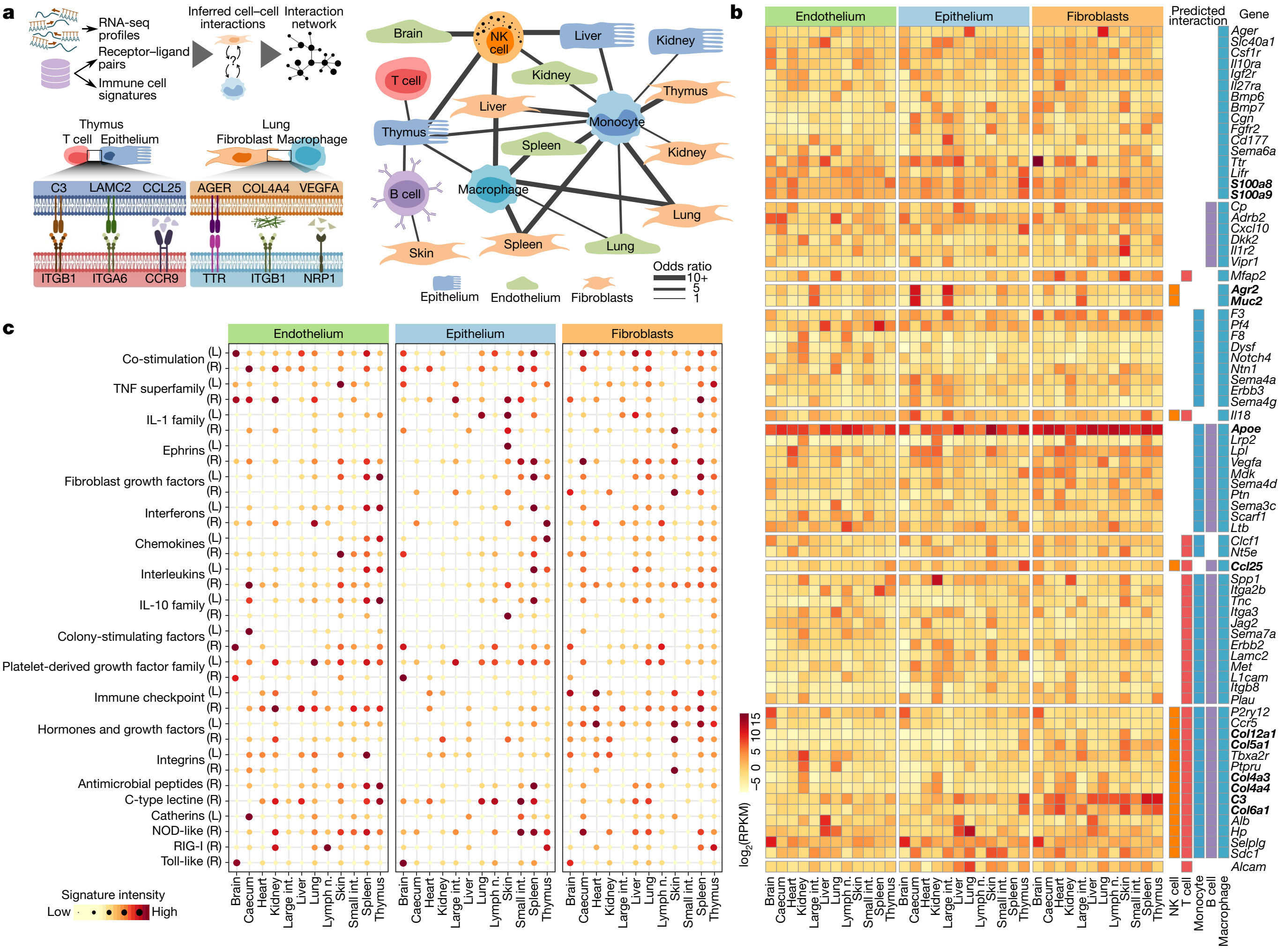


Fig. 2 | Gene expression of structural cells predicts cell-type-specific and organ-specific crosstalk with haematopoietic immune cells. a, Network of potential cell–cell interactions between structural cells and haematopoietic immune cells inferred from gene expression of known receptor–ligand pairs. NK cell, natural killer cell. **b**, Expression of receptors and ligands in structural cells, annotated with the cell–cell interactions that they may mediate (genes discussed in the text are in bold). RPKM, reads per kilobase of transcript per million mapped reads. **c**, Gene signatures of receptors (R) and ligands (L) in structural cells. Sample size (all panels): $n = 3$ independent biological replicates.

cells and haematopoietic cells under homeostasis (Fig. 2a, Extended Data Fig. 4a). Differences across cell types and organs were driven by the characteristic expression patterns of cell-surface proteins and secreted factors in structural cells (Fig. 2b). For example, strong expression of collagens (*Col4a3*, *Col4a4*, *Col5a1*, *Col6a1* and *Col12a1*) and of complement component 3 (*C3*) in fibroblasts is expected to enhance their ability to interact with haematopoietic cells; high levels of *Muc2* and *Arg2* in the digestive tract fosters interactions with macrophages and natural killer cells; and *Ccl25* expression in thymus epithelium contributes to the maturation of T cells. Shared across all types of structural cells, we observed high expression of the inflammatory mediators *Apoe*, *S100a8* and *S100a9*.

Second, we quantified the aggregated activity of various immune gene modules, which were manually curated to capture important components of the immune system (Extended Data Fig. 4b, Supplementary Table 4). We observed widespread activity of these immune gene modules in structural cells, with highly cell-type-specific and organ-specific patterns (Fig. 2c). Next, we sought to validate the unexpectedly strong and diverse immune gene activity of structural cells in a second dataset. To that end, we obtained single-cell transcriptome profiles of mouse tissues from the Tabula Muris resource[20], and we identified endothelium, epithelium and fibroblasts bioinformatically on the basis of the expression of marker genes. Although structural cells were not well covered in this dataset (Extended Data Fig. 5a, Supplementary Table 5), these profiles were sufficient to independently confirm strong and organ-specific activity of immune genes in structural cells (Extended Data Fig. 5b–d). Together, these data reveal widespread activity of immune genes and regulatory modules in structural cells, which are expected to mediate cell-type-specific and organ-specific interactions with haematopoietic immune cells.

Regulatory networks in structural cells

To uncover the regulatory basis of immune gene activity in structural cells, we combined our RNA-seq data with ATAC-seq profiles of chromatin accessibility and ChIPmentation maps for the promoter and enhancer-linked H3K4me2 mark, and we compared gene-regulatory networks across cell types and organs (Fig. 3a).

We observed extensive cell-type-specific and organ-specific chromatin regulation at immune gene loci (Fig. 3b). For example, the immune-regulatory transcription factors *Stat5a* and *Stat5b* showed characteristic patterns of chromatin accessibility; *Tlr9* and *Osm* were characterized by organ-specific heterogeneity in promoter and enhancer regions; and the promoter of the renal cell adhesion molecule

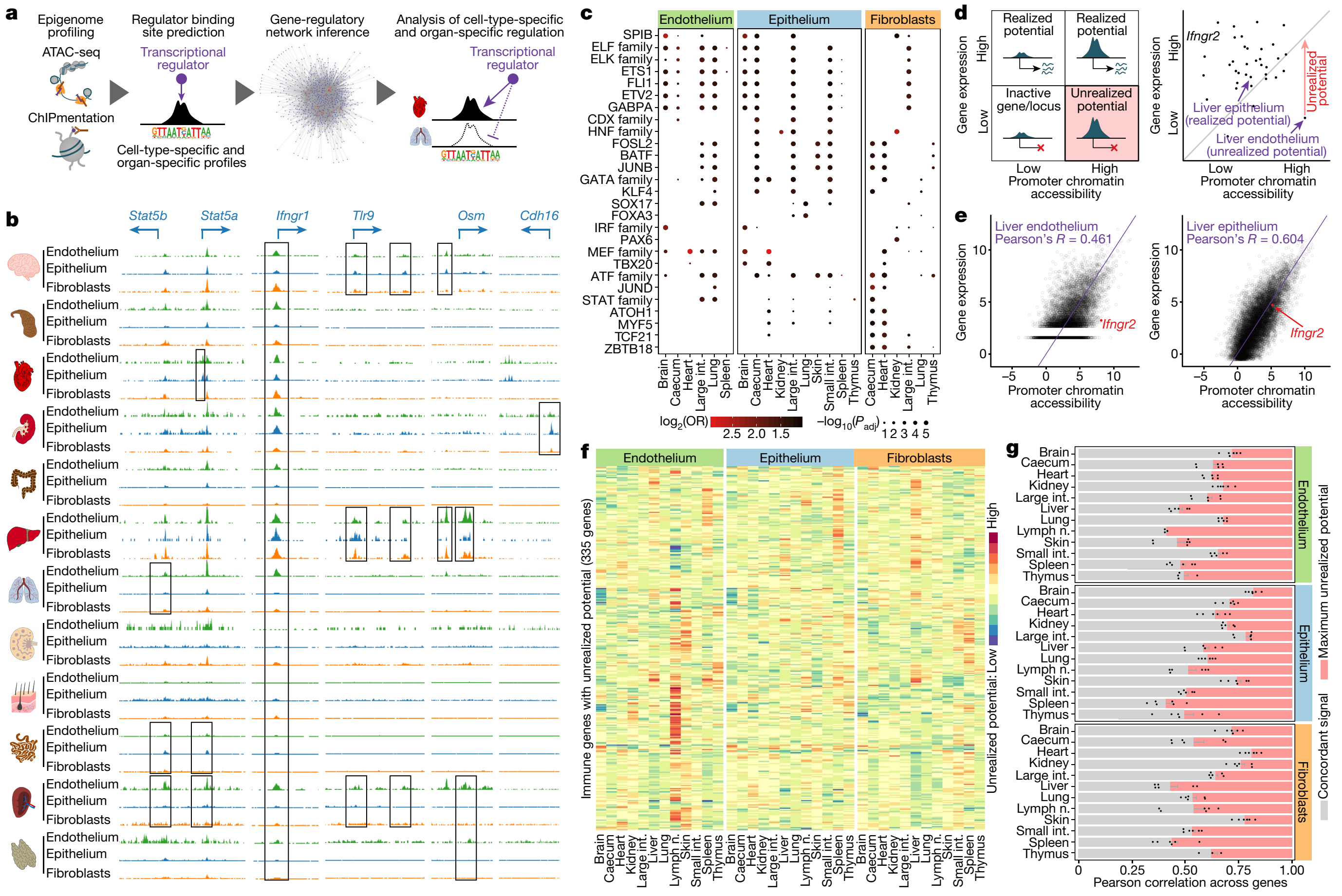


Fig. 3 | Structural cells implement characteristic gene-regulatory networks and an epigenetic potential for immune gene activation. **a**, Schematic outline of the gene-regulatory network analysis. **b**, ATAC-seq signal tracks (average across replicates) for selected genomic regions. Differences between cell types and organs are highlighted by black boxes. **c**, Motif enrichment for transcriptional regulators in cell-type-specific and organ-specific chromatin marker peaks (one-sided hypergeometric test with multiple-testing correction). OR, odds ratio; P_{adj}, adjusted *P* value. **d**, Schematic outline (left) and a concrete example (right) of the epigenetic potential, based on the comparison of chromatin accessibility (ATAC-seq) in promoter regions with matched gene expression (RNA-seq). **e**, Scatterplot showing the correlation between promoter chromatin accessibility and gene expression across all genes in liver endothelium and epithelium, with *Ifngr2* highlighted by the red dot. **f**, Immune genes with unrealized epigenetic potential across cell types and organs. **g**, Pearson correlation between promoter chromatin accessibility and gene expression across cell types and organs (mean and s.e.m. across pairwise correlations; red bars indicate the maximum scope for unrealized epigenetic potential). Sample size: ATAC-seq *n* = 2 (**b**–**g**), RNA-seq *n* = 3 (**c**–**g**) independent biological replicates.

Cdh16 was exclusively open in kidney. By contrast, a subset of crucial immune genes (exemplified by *Ifngr1*) showed high chromatin accessibility in most samples, indicative of a shared core of immune regulation in structural cells. Consistent with our RNA-seq analysis, the chromatin profiles were globally more similar between different types of structural cells in one organ than among the same cell type across organs (Extended Data Fig. 6a, b).

We inferred a gene-regulatory network of structural cells by connecting transcription factors to their target genes, based on predicted binding sites with open chromatin in the respective cells (Extended Data Fig. 6c). Many key regulators of transcription in structural cells showed cell-type-specific and organ-specific activity (Fig. 3c, Extended Data Fig. 6d, Supplementary Table 6). For example, ATF, ELK, ETS and JUND were most active in lung endothelium; KLF and CDX in digestive tract epithelium; and HNF in kidney fibroblasts and in epithelium of caecum, large intestine and small intestine. We also identified groups of transcription factors that were ubiquitously active in structural cells, such as ELF1, ELF3, ETS1, FLI1 and GATA2 (Extended Data Fig. 6e), which may constitute a shared regulatory basis of immune gene activity in structural cells. Our inferred gene-regulatory networks support a model in which constitutively active regulators establish a shared core of immune functions, while additional factors contribute cell-type-specific and organ-specific adaptions.

Epigenetic potential for gene activation

Epigenome profiles not only capture the current regulatory states of cells, but also reflect their future potential to respond to various stimuli[21]. We thus hypothesized that structural cells may be epigenetically primed for immune gene activation, which would pre-program them for a rapid response to immunological challenges.

To assess the epigenetically encoded immune potential of structural cells, we quantified the chromatin accessibility of each gene promoter and compared it to the expression level of the corresponding gene. We then scanned for genes with low expression but high promoter accessibility, indicative of an unrealized potential for increased expression (Fig. 3d, e, Extended Data Fig. 7a). For example, chromatin accessibility at the *Ifngr2* promoter was high in liver endothelium whereas gene expression was low, which suggests that *Ifngr2* has unrealized potential for upregulation without the need to increase promoter accessibility. By contrast, accessibility of the *Ifngr2* promoter in liver epithelium was highly consistent with its gene expression, thus constituting a case of realized potential.

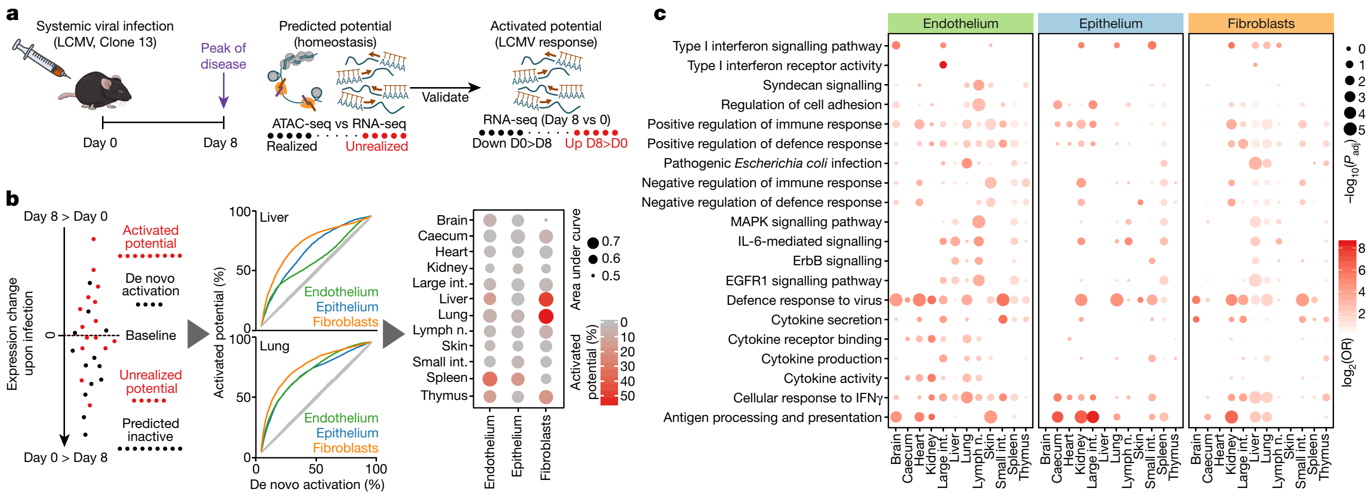


Fig. 4 | Systemic viral infection activates the immunological potential of structural cells in vivo. a, Schematic outline of the LCMV infection analysis. **b**, Comparison of the changes in gene expression after LCMV infection (day 8) to the epigenetic potential observed under homeostatic conditions (day 0), using a threshold of zero for differential gene expression (left) or a variable threshold analogous to a receiver operating characteristic curve (centre). The area under the curve is interpreted as a measure of the predictiveness of the epigenetic potential for LCMV-induced gene activation, plotted together with the percentage of upregulated genes that carry unrealized epigenetic potential (right). **c**, Enrichment of immune-related gene sets among the LCMV-induced genes (two-sided Fisher's exact test with multiple-testing correction). Sample size (all panels): $n = 3$ independent biological replicates.

The genes with unrealized epigenetic potential in structural cells were enriched for immune functions (Fig. 3f, Extended Data Fig. 7b, Supplementary Tables 7 and 8), consistent with our initial hypothesis. In total, we identified 1,665 genes that fulfilled our definition of unrealized epigenetic potential, of which 335 genes were annotated with at least one immunological term (odds ratio 1.37, $P < 10^{-5}$, Fisher's exact test).

To quantify and compare the epigenetic potential across cell types and organs (Fig. 3g), we exploited that genes with strong unrealized potential are outliers when plotting gene expression against promoter accessibility, which results in a reduced correlation between the two data types (Extended Data Fig. 7a). The highest correlation between gene expression and promoter accessibility was observed in brain, caecum, heart, kidney, large intestine and skin, leaving comparatively less room for unrealized epigenetic potential. By contrast, the correlation was notably lower in liver, lymph node, spleen and thymus, which suggests that structural cells in these organs harbour a more pronounced epigenetic potential for gene activation in response to various stimuli.

Our integrative analysis of chromatin accessibility and gene expression thus identified an epigenetically encoded potential for immune gene activation in structural cells. This epigenetic potential is expected to facilitate the rapid response to immunological challenges in a cell-type-specific and organ-specific manner.

Immune genes induced by viral infection

We functionally evaluated the epigenetic potential for immune gene activation by challenging mice with a systemic viral infection model. We infected mice with lymphocytic choriomeningitis virus (LCMV) and collected samples from 12 organs on day 8 after infection (Fig. 4a). We characterized these samples by flow cytometry (Extended Data Fig. 8a, b) and by RNA-seq analysis of sort-purified structural cells (Supplementary Table 1).

LCMV infection resulted in changes of structural cell composition in most organs (Extended Data Fig. 8c), and we observed differential gene expression in a cell-type-specific and organ-specific manner (Extended Data Fig. 9a, Supplementary Table 9). The transcriptional response was globally more similar among structural cells of all three types within the same organ than between structural cells of a given type across organs (Extended Data Fig. 9b), consistent with the patterns of similarity that we observed under homeostatic conditions.

We compared the differential gene expression after LCMV infection with the corresponding epigenetic potential of the genes under homeostatic conditions, and genes with unrealized potential were indeed overrepresented among the LCMV-induced genes (Fig. 4b, left). To quantify how well the epigenetic potential predicts LCMV-induced gene activation, we plotted the percentage of genes with realized potential against the percentage of genes that underwent de novo activation (Fig. 4b, centre), in analogy with receiver operating characteristic (ROC) curves. We detected the strongest association in liver, lung and spleen (Fig. 4b, right). These organ-specific differences showed no clear correlation with differences in viral load (Extended Data Fig. 9c) and appear to constitute intrinsic regulatory differences between cell types and organs (Extended Data Fig. 9d).

The genes that were upregulated in response to systemic LCMV infection showed strong enrichment for immune functions (Fig. 4c), including 'positive regulation of immune response', 'defence response to virus', 'cellular response to IFNγ' and 'antigen processing and presentation'. We also observed widespread upregulation of interferon-induced as well as interferon-stimulated genes and of key transcriptional regulators in the interferon pathway, which indicates a strong interferon response to LCMV infection in structural cells (Supplementary Table 9).

Finally, we inferred receptor–ligand interactions between structural cells and haematopoietic immune cells, using the transcriptome data of structural cells upon LCMV infection (Extended Data Fig. 9e, f, Supplementary Table 10). We observed an increase in the strength and scope of predicted cell–cell interactions after LCMV infection as compared to homeostatic conditions, largely driven by upregulated gene expression levels for receptors and ligands in structural cells, including *B2m*, *Cd74*, *Cd47*, *Cxcl10*, *Sdc1*, *Sdc4*, *Tnfrsf1a* and *Vcam1*.

Systemic LCMV infection thus triggered widespread activation of immune genes that were lowly expressed but epigenetically poised under homeostatic conditions. These results support our model of an epigenetically encoded potential for immune gene activation in structural cells in the context of viral infection.

Cytokine response of structural cells

To characterize the effects of individual cytokines that may contribute to the response to LCMV infection, we administered six recombinant cytokines in mice (IFNα, IFNγ, IL-3, IL-6, TGFβ and TNF). These cytokines

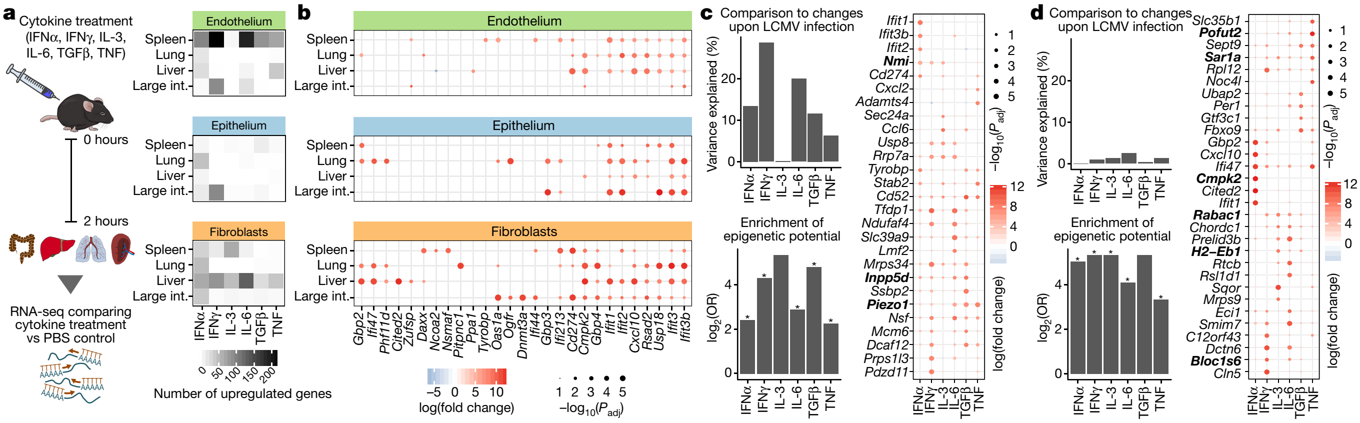


Fig. 5 | Cytokine treatment induces cell-type-specific and organ-specific changes in structural cells in vivo. a, Schematic outline of the cytokine treatment experiments (left) and number of genes upregulated in each experiment (right). **b**, Genes upregulated in response to treatment with IFNα. **c**, **d**, Cytokine-induced changes in spleen endothelium (**c**) and liver fibroblasts (**d**), showing the percentage of LCMV-induced changes that are recapitulated by cytokine treatment (top left), enrichment for genes with unrealized epigenetic potential among the cytokine-induced genes (bottom left), and genes upregulated upon cytokine treatment (genes discussed in the text are in bold). Significant enrichments (two-sided Fisher's exact test, adjusted $P < 0.05$) are labelled with an asterisk. Differential expression is based on a linear model (two-sided test) with multiple-testing correction (**b**–**d**). Sample size (all panels): $n = 3$ independent biological replicates.

were selected based on our dataset and published results[22], and IL-3 was included as a control with no known role in LCMV infection. To focus on the immediate effects of cytokine signalling, structural cells were sort-purified two hours after cytokine injection and subjected to RNA-seq profiling (Fig. 5a, Supplementary Table 1). Four organs that responded strongly to LCMV infection were included in the analysis (large intestine, liver, lung and spleen).

We compared the structural cell transcriptomes between mice that were cytokine-treated and corresponding mock-treated controls, which uncovered various cytokine-induced transcriptional changes. Treatment with IFNα had the strongest effect, inducing known interferon target genes and several genes associated with antiviral immunity (Fig. 5b, Extended Data Fig. 10, Supplementary Table 11). We further observed widespread cell-type-specific and organ-specific differences (Extended Data Fig. 11a), which were not solely due to differential expression of individual cell surface receptors (Extended Data Fig. 11b), but appear to reflect more general differences in immune gene regulation.

In spleen endothelium (highlighted here because of its strong transcriptional response to LCMV infection), treatment with IFNγ and IL-6 explained a sizable proportion of the LCMV-induced changes (Fig. 5c, top left). The cytokine-induced genes included known interferon target genes, transcriptional regulators such as *Nmi* and *Inpp5d* (which encodes SHIP1), and the pro-inflammatory gene *Piezo1* (Fig. 5c, right, Supplementary Table 11). For five of the six cytokines (not including IL-3), the upregulated genes showed significant overlap with the genes that carried unrealized epigenetic potential under homeostatic conditions (Fig. 5c, bottom left). These cytokines thus triggered similar aspects of the epigenetic potential in spleen endothelium as observed for systemic LCMV infection.

The results were qualitatively different in liver fibroblasts. We observed little overlap between the transcriptional response to the cytokines and to LCMV infection (Fig. 5d, top left), while there was still a strong association with the epigenetic potential (Fig. 5d, bottom left). Cytokine administration induced genes involved in metabolic processes (*Cmpk2*, *Pofut2*) as well as regulators of vesicle or membrane trafficking and antigen presentation (*Bloc1s6*, *H2-Eb1*, *Rabac1* and *Sar1a*) (Fig. 5d, right, Supplementary Table 11). Thus, the administration of cytokines triggered different aspects of the epigenetic potential in liver fibroblasts compared with LCMV infection.

In summary, structural cells responded in cell-type-specific and organ-specific ways to in vivo stimulation with individual cytokines, which allowed us to functionally dissect the effects of systemic LCMV infection and provides further validation of the observed epigenetic potential for immune gene activity in structural cells.

Discussion

Structural cells, including endothelium, epithelium and fibroblasts, are important yet underappreciated contributors to mammalian immune responses. Here, we systematically investigated immune gene regulation in these non-haematopoietic cell types. To that end, we applied multi-omics profiling and integrative bioinformatic analysis to three types of structural cells purified from twelve different organs of the mouse. We observed unexpectedly strong and densely regulated expression of immune genes, both under homeostatic conditions and in response to immunological challenges (systemic viral infection with LCMV, in vivo cytokine treatment).

Immunologists tend to consider structural cells mainly for their barrier function (epithelium, endothelium) and their role as connective tissue (fibroblasts), although important research has identified much more diverse roles of structural cells in mammalian immunity[23–31]. However, comparative multi-organ investigations of structural cells have been lacking. We sought to close this gap in our understanding of mammalian immunity by evaluating the immune regulation of structural cells in a systematic, genome-wide and organism-scale way. We identified three main lines of evidence that highlight the immune-regulatory potential of structural cells.

First, we observed extensive cell-type-specific and organ-specific regulation of genes that influence the capabilities of structural cells to engage in predicted interactions with haematopoietic immune cells. On the basis of our transcriptome data, we inferred an initial network of potential cell–cell interactions between structural cells and haematopoietic cells. It will be an important future goal to dissect these cell–cell interactions and to untangle the precise chain of command in the immunological communication of structural cells and haematopoietic cells.

Second, our genome-wide analysis of chromatin accessibility in structural cells uncovered not only a regulatory basis of their immune functions, but also an initial assessment of the transcriptional regulators that confer cellular identity to endothelial cells, epithelial cells and fibroblasts across 12 organs. Future studies could pursue genetic manipulation of these candidate regulators in a cell-type-specific and organ-specific way, investigating the effect on immune-regulatory functions, epigenetic landscapes and cellular identity.

Third, our integrative analysis of gene expression and chromatin accessibility identified an epigenetically encoded immune potential in structural cells, constituted by genes that were lowly expressed under homeostatic conditions but epigenetically poised for much higher expression. These genes were enriched for immune functions, and they were preferentially upregulated in response to LCMV infection and cytokine administration in vivo. We thus conclude that structural cells are epigenetically pre-programmed for a swift response to a variety of immunological challenges. It will be interesting to explore how the epigenetic potential of structural cells responds to other stimuli and whether it can be modulated for therapeutic purposes, for example in the context of autoimmune diseases or the tumour microenvironment.

In conclusion, our study provides a comprehensive characterization of immune gene regulation in structural cells, and an initial step towards the systematic, organism-scale dissection of immune functions beyond haematopoietic cells. To emphasize the importance of structural cells for mammalian immunity, we tentatively propose the term 'structural immunity' for the study of immune functions in the non-haematopoietic, structural cell populations of the body. We see our study and large-scale dataset as a starting point, reference atlas, and a collection of hypotheses for systematic as well as mechanistic explorations in this emerging area of research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2424-4.

1. Kotas, M. E. & Medzhitov, R. Homeostasis, inflammation, and disease susceptibility. *Cell* **160**, 816–827 (2015).
2. Abbas, A. K., Lichtman, A. H. & Pillai, S. *Cellular and Molecular Immunology* 9th edn (Elsevier, 2017).
3. Buckley, C. D., Barone, F., Nayar, S., Bénézech, C. & Caamaño, J. Stromal cells in chronic inflammation and tertiary lymphoid organ formation. *Annu. Rev. Immunol*. **33**, 715–745 (2015).
4. Pober, J. S. & Sessa, W. C. Evolving functions of endothelial cells in inflammation. *Nat. Rev. Immunol*. **7**, 803–815 (2007).
5. Schleimer, R. P., Kato, A., Kern, R., Kuperman, D. & Avila, P. C. Epithelium: at the interface of innate and adaptive immune responses. *J. Allergy Clin. Immunol*. **120**, 1279–1284 (2007).
6. Pawlina, W. & Ross, M. H. *Histology: a Text and Atlas, International Edition: with Correlated Cell and Molecular Biology* (Wolters Kluwer Law & Business, 2019).
7. Humbert, M., Hugues, S. & Dubrot, J. Shaping of peripheral T cell responses by lymphatic endothelial cells. *Front. Immunol*. **7**, 684 (2017).
8. Malhotra, D., Fletcher, A. L. & Turley, S. J. Stromal and hematopoietic cells in secondary lymphoid organs: partners in immunity. *Immunol. Rev*. **251**, 160–176 (2013).
9. Nowarski, R., Jackson, R. & Flavell, R. A. The stromal intervention: Regulation of immunity and inflammation at the epithelial-mesenchymal barrier. *Cell* **168**, 362–375 (2017).
10. Perez-Shibayama, C., Gil-Cruz, C. & Ludewig, B. Fibroblastic reticular cells at the nexus of innate and adaptive immune responses. *Immunol. Rev*. **289**, 31–41 (2019).
11. Turley, S. J., Cremasco, V. & Astarita, J. L. Immunological hallmarks of stromal cells in the tumour microenvironment. *Nat. Rev. Immunol*. **15**, 669–682 (2015).
12. Davis, M. M., Tato, C. M. & Furman, D. Systems immunology: just getting started. *Nat. Immunol*. **18**, 725–732 (2017).
13. Villani, A. C., Sarkizova, S. & Hacohen, N. Systems immunology: learning the rules of the immune system. *Annu. Rev. Immunol*. **36**, 813–842 (2018).
14. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
15. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
16. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods* **12**, 963–965 (2015).
17. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222–228 (2014).
18. Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
19. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
20. The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
21. Bock, C. et al. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
22. Baazim, H. et al. CD8$^+$ T cells induce cachexia during chronic viral infection. *Nat. Immunol*. **20**, 701–710 (2019).
23. Guerrero-Juarez, C. F. et al. Single-cell analysis reveals fibroblast heterogeneity and myeloid-derived adipocyte progenitors in murine skin wounds. *Nat. Commun*. **10**, 650 (2019).
24. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386 (2018).
25. Mizoguchi, F. et al. Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat. Commun*. **9**, 789 (2018).
26. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
27. Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).
28. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
29. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
30. Rodda, L. B. et al. Single-cell RNA sequencing of lymph node stromal cells reveals niche-associated heterogeneity. *Immunity* **48**, 1014–1028 (2018).
31. Yan, K. S. et al. Non-equivalence of Wnt and R-spondin ligands during Lgr5$^+$ intestinal stem-cell self-renewal. *Nature* **545**, 238–242 (2017).

Methods

Mice

C57BL/6J mice were bred and maintained under specific pathogen free conditions at the Institute of Molecular Biotechnology (IMBA) of the Austrian Academy of Sciences in Vienna (Austria). In vivo experiments (systemic viral infection, cytokine treatments) were performed under specific pathogen-free conditions at the Anna Spiegel Research Building of the Medical University of Vienna (Austria). Age-matched male mice (8 to 13 weeks old) were used in all experiments. For the characterization of structural cells under homeostatic conditions, mice were killed without any previous treatment. For the systemic viral infection experiments, mice were intravenously infected with 2×10^6 focus-forming units of LCMV strain clone 13[32,33] and killed on day 8 after infection. For the cytokine treatment experiments, mice were intravenously injected with 100 $\mu g\ kg^{-1}$ of the following recombinant cytokines: IFNα, IFNγ, IL-3, IL-6, TGFβ or TNF (all from BioLegend) and killed 2 h after cytokine injection. All mouse experiments were performed in individually ventilated cages according to the respective animal experiment licenses (BMWFW-66.009/0199-WF/V/3v/2015 and BMWFW-66.009/0361-WF/V/3b/2017) approved by the institutional ethical committees and the institutional guidelines at the Department for Biomedical Research of the Medical University of Vienna. Samples numbers are listed in the figure legends. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to treatment status during experiments and outcome assessment.

Standardized sample collection and organ dissociation

Different surface markers and sorting schemes were previously used to purify endothelium, epithelium and fibroblasts in individual organs, whereas our study required standardized cell purification across organs. We therefore tested several surface markers across organs and optimized a sorting scheme that produced consistent results for all 12 investigated organs, while excluding cell types that were detectable only in one or a few tissues (most notably lymphatic endothelial cells). The experimental workflow followed the recommendations of the Immunological Genome Project[34] regarding sample collection schedule, antibody staining and sample pooling. At least three same-sex littermate mice were pooled for each biological replicate for homeostatic conditions and for LCMV infection, respectively. For the in vivo cytokine treatments, we used individual mice as biological replicates to reduce the total number of mice. Standardized organ harvesting and dissociation protocols were established to obtain single-cell suspension for subsequent cell purification by fluorescence activated cell sorting (FACS). The same digestion solution was used for all organs, to avoid organ-specific technical confounders. The workflow is described in detail below.

Brain. After decapitation, the skull was cut longitudinally with scissors, and the cranium was opened with tweezers. Both brain hemispheres were carefully collected and placed into cell culture dishes containing cold PBS supplemented with 0.1% BSA (PBS + BSA). White matter was manually removed, the tissue shredded with scissors and added to a 50 ml tube containing 15 ml cold Accumax (Sigma-Aldrich) digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer (Sigma-Aldrich) followed by centrifugation at 300*g* for 5 min at 4 °C. Myelin was removed by using density gradient centrifugation. Cells were recovered at the interface between an 80% Percoll layer and a 30% Percoll layer and washed in PBS + BSA to remove excess Percoll.

Caecum, large intestine and small intestine. Luminal contents were removed. The organs were cut longitudinally with scissors, rinsed several times in PBS + BSA to remove mucus, then cut into 0.5 cm pieces and placed in 50 ml tubes containing 20 ml pre-warmed (37 °C) RPMI containing 10% FCS and 5 mM EDTA (RPMI + FCS + EDTA). Samples were incubated for 25 min at 37 °C while shaking at 200 rpm. Supernatant was collected, samples were resuspended once again in 20 ml pre-warmed RPMI + FCS + EDTA, and incubated for 25 min at 37 °C in a shaking incubator at 200 rpm. These wash steps were performed to dissociate epithelial cells. After the second incubation, supernatants were collected and combined, followed by digestion of the samples in 15 ml cold Accumax for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer. Organ homogenates were combined with the epithelial cell fractions, filtered twice through a 100-µm cell strainer and washed in cold PBS + BSA.

Heart, kidney, lung, spleen and thymus. Organs were rinsed with cold PBS + BSA and shredded with scissors. Tissue fragments were placed into 50 ml tubes containing 20 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered through a 100-µm cell strainer. After centrifugation, 2 ml cold ACK lysis buffer (Thermo Fisher Scientific) was added for 3 min to lyse red blood cells and the reaction stopped by adding 20 ml of cold PBS + BSA. Supernatants were filtered twice through a 100-µm cell strainer and washed once to remove residual ACK lysis buffer.

Liver. Three lobes were removed, rinsed with cold PBS + BSA, and shredded with scissors. Tissue fragments were placed into a 50 ml tube containing 20 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered through a 100-µm cell strainer. Hepatocytes were removed using density gradient centrifugation. Cells recovered at the interface between an 80% Percoll layer and a 30% Percoll layer were washed in PBS + BSA to remove excess Percoll.

Lymph nodes. Cervical, axillary and inguinal lymph nodes were combined, carefully pinched with tweezers, and rinsed several times with cold PBS + BSA to release haematopoietic cells. Tissue fragments were placed into a 50 ml tube containing 10 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered twice through a 100-µm cell strainer.

Skin. Ears were obtained at the base, and the subcutaneous fat layer scrapped off with a scalpel. Tissue fragments were then shredded with scissors, placed into 50 ml tubes containing 15 ml of Accumax digestion solution, and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered twice through a 100-µm cell strainer.

Organ-specific sample collection and organ dissociation

Large intestine. The organ was removed and processed as described[35]. In brief, luminal contents were removed, the large intestine cut longitudinally with scissors, rinsed several times in PBS + BSA to remove mucus, then cut into 0.5 cm pieces and placed in 50 ml tubes containing 20 ml pre-warmed (37 °C) RPMI containing 10% FCS and 5 mM EDTA (RPMI + FCS + EDTA). Samples were incubated for 40 min at 37 °C while shaking at 200 rpm. Supernatant was collected, samples were resuspended once again in 20 ml pre-warmed RPMI + FCS + EDTA and incubated for 20 min at 37 °C while shaking at 200 rpm. These wash steps were performed to dissociate epithelial cells. After the second incubation, supernatants were collected and combined, followed by incubating samples in RPMI containing 10% FCS and 15 mM HEPES (RPMI + FCS + HEPES) for 10 min at room temperature. Supernatant was discarded and samples digested in RPMI + FCS + HEPES containing 100 U ml^{-1} collagenase from *Clostridium histolyticum* (Sigma) for 1 h at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer. Organ homogenates were combined with the epithelial cell fractions, filtered twice through 100-μm cell strainers and washed in cold PBS + BSA.

Lung. The organ was removed and processed as described[36]. In brief, the organ was cut into 0.5-cm pieces and placed in gentleMACS C tubes (Miltenyi) containing 160 U ml^{-1} collagenase type 1 (Gibco) and 12 U ml^{-1} DNase 1 (Sigma) in RPMI containing 5% FCS, and dissociated using a gentleMACS Dissociator (Miltenyi; program m_lung_01). After incubation at 37 °C for 35 min while shaking at 170 rpm, digested samples were homogenized using the gentleMACS Dissociator (program m_lung_02). Subsequently, cell suspensions were filtered through 70-μm cell strainers and centrifuged for 5 min at 4 °C at 300*g*. After centrifugation, 1 ml cold ACK lysis buffer (Thermo Fisher) was added for 5 min to lyse red blood cells and the reaction stopped by adding 20 ml of cold PBS + BSA. Supernatants were filtered twice through 100-μm cell strainer and washed once to remove residual ACK lysis buffer.

Liver. Mice were anaesthetized (ketamine:xylazine 1:3, 0.1 ml per 10 g mouse; Vetoquinol) before cannulation of the liver and dissociation using a two-step perfusion protocol[37]. In brief, the liver was perfused first with 20 ml HBSS (Gibco) containing 0.5 mM EGTA (Sigma) and then with 20 ml L15 medium (Gibco) containing 40 mg l^{-1} liberase (Roche) at a rate of 5 ml min^{-1}. Next, the liver was removed, placed in a Petri dish with 10 ml of the same liberase-containing medium, followed by removal of the liver capsule. Hepatocytes were removed from the resulting cell suspension using density gradient centrifugation. Cells recovered at the interface between an 80% Percoll layer and a 30% Percoll layer were washed in PBS + BSA to remove excess Percoll.

Flow cytometry and FACS

Single-cell suspensions were washed once with PBS containing 0.1% BSA and 5 mM EDTA (PBS + BSA + EDTA). Cells were then incubated with anti-CD16/CD32 (clone 93, BioLegend) to prevent nonspecific binding. Single-cell suspensions were then stained with different combinations of antibodies against CD45 (PerCP-Cy5.5, clone 30-F11), TER-119 (PerCP-Cy5.5, clone TER-119), GP38 (PE, clone 8.1.1), EpCAM (Pe-Cy7, clone 8.8), LTβP (APC, clone 5G11), CD31 (FITC, clone MEC13.3), CD90.2 (AF700, clone 30-H12), CD106 (AF647, clone 429), CD144 (BV421, clone BV13), CD324 (APC-Cy7, clone DECMA-1) (all from BioLegend), CD140a (BV605, clone APA5, BD Bioscience), MAdCAM1 (BV421, clone MECA-367, BD Bioscience) and LYVE1 (eFlour 660, clone ALY7, Thermo Fisher Scientific) for 30 min at 4 °C, followed by two washes with PBS + BSA + EDTA. Dead cells were stained by adding either Zombie Red Fixable Viability Dye or Zombie Aqua Fixable Viability Dye (both from BioLegend) immediately before flow cytometry characterization or cell sorting. For flow cytometry, cells were acquired with an LSRFortessa (BD Biosciences) cell analyser using the outlined gating strategies (Extended Data Figs. 1a, 2a, 8a). For FACS, cells were sort-purified with a MoFlo Astrios (Beckman Coulter) or SH800 (Sony) using the outlined gating strategies (Extended Data Figs. 1a, 8a). Data analysis was performed using the FlowJo software (v.10.5.3, Tree Star).

Transcriptome profiling by Smart-seq2

Smart-seq2 was performed as previously described[14], starting from low-input bulk samples. In each experiment, a maximum of 200 cells were sort-purified and deposited in 96-well plates containing 4 μl lysis buffer (1:20 solution of RNase Inhibitor (Clontech) in 0.2% (v/v) Triton X-100 (Sigma-Aldrich)), spun down and immediately frozen at −80°C. Reverse transcription was performed using SuperScript II (Invitrogen) followed by PCR amplification with KAPA HiFi HotStart Ready Mix (Kapa Biosystems). cDNA amplification was followed by two rounds of SPRI (Beckman Coulter) purification, and cDNA concentration was measured with a Qubit fluorometer (Life Technologies). Library preparation was conducted on 1 ng of cDNA using the Nextera XT DNA Sample Preparation Kit (Illumina), followed by SPRI (Beckman Coulter) size selection. Libraries were sequenced by the Biomedical Sequencing Facility at CeMM using the Illumina HiSeq 3000/4000 platform and the 50-bp single-end configuration. Transcriptome profiling by Smart-seq2 was done in three biologically independent experiments. Sequencing statistics are provided in Supplementary Table 1.

Chromatin accessibility mapping by ATAC-seq

ATAC-seq was performed as previously described[15,38], with minor adaptations. In each experiment, a maximum of 50,000 sort-purified cells were collected at 300*g* for 5 min at 4 °C. After centrifugation, the pellet was carefully resuspended in the transposase reaction mix (12.5 μl 2xTD buffer, 2 μl TDE1 (Illumina), 10.25 μl nuclease-free water and 0.25 μl 1% digitonin (Promega)) for 30 min at 37 °C. Following DNA purification with the MinElute kit eluting in 11 μl, 1 μl of eluted DNA was used in a quantitative PCR (qPCR) reaction to estimate the optimum number of amplification cycles. The remaining 10 μl of each library were amplified for the number of cycles corresponding to the C_q value (that is, the cycle number at which fluorescence has increased above background levels) from the qPCR. Library amplification was followed by SPRI (Beckman Coulter) size selection to exclude fragments larger than 1,200 bp. DNA concentration was measured with a Qubit fluorometer (Life Technologies). Library amplification was performed using custom Nextera primers[15]. Libraries were sequenced by the Biomedical Sequencing Facility at CeMM using the Illumina HiSeq 3000/4000 platform and the 50-bp single-end configuration. Chromatin accessibility mapping by ATAC-seq was done in two biologically independent experiments. Sequencing statistics are provided in Supplementary Table 1.

Epigenome mapping by ChIPmentation for H3K4me2

ChIPmentation was performed as previously described[16,39], with minor adaptions. In each experiment, a maximum of 50,000 sort-purified cells were washed once with PBS and fixed with 1% paraformaldehyde for 10 min at room temperature. Glycine (0.125 M final concentration) was added to stop the reaction. Cells were collected at 500*g* for 10 min at 4 °C (subsequent work was performed on ice and used cooled buffers and solutions unless otherwise specified) and washed once with ice-cold PBS supplemented with 1 mM phenylmethyl sulfonyl fluoride (PMSF). After centrifugation, cells were lysed in sonication buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.25% SDS, 1× protease inhibitors (Sigma-Aldrich) and 1 mM PMSF) and sonicated (Covaris S220) for 30 min in a microTUBE until the size of most fragments was in the range of 200 to 700 bp. Following sonication, the lysate was adjusted to RIPA buffer conditions (final concentration: 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, 1× protease inhibitors (Sigma-Aldrich) and 1 mM PMSF).

For each immunoprecipitation, 10 μl magnetic Protein A (Life Technologies) were washed twice and resuspended in PBS supplemented with 0.1% BSA. One microgram of antibody recognizing H3K4me2 (Sigma-Aldrich, clone AW30) was added and bound to the beads by rotating overnight at 4 °C. Beads were added to the sonicated lysate and incubated for 2 h at 4 °C on a rotator followed by washing the beads once with RIPA low-salt buffer (10 mM Tris-HCl, 150 mM NaCl, 0.1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100 and 1 mM EDTA), once with RIPA high-salt buffer (10 mM Tris-HCl, 500 mM NaCl, 0.1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100 and 1 mM EDTA), once with RIPA lithium-chloride buffer (10 mM Tris-HCl, 250 mM LiCl, 0.5% IGEPAL CA-630, 0.5% sodium deoxycholate and 1 mM EDTA) and once with Tris-Cl pH 8. Bead-bound chromatin was then resuspended in tagmentation mix (5 μl 5×TD buffer, 1 μl TDE1 (Illumina), 19 μl nuclease-free water) and incubated for 10 min at 37 °C.

After tagmentation, the beads were washed once with RIPA and once with cold Tris-Cl pH 8. Bead bound tagmented chromatin was resuspended in 10.5 μl 20 mM EDTA and incubated for 30 min at 50 °C. Then, 10.5 µl 20 mM $MgCl_2$ as well as 25 µl 2× KAPA HiFi HotStart Ready Mix (Kapa Biosystems), pre-activated by incubation at 98 °C for 45 s, were added and incubated for 5 min at 72 °C, followed by incubation for 10 min at 95 °C. Beads were magnetized and 2 µl of each library were amplified in a 10 µl qPCR reaction containing 0.8 mM primers, SYBR Green and 5 µl Kapa HiFi HotStart ReadyMix to estimate the optimum number of enrichment cycles, using the following program: 72 °C for 5 min, 98 °C for 30 s, 24 cycles of 98 °C for 10 s, 63 °C for 30 s, 72 °C for 30 s, and a final elongation at 72 °C for 1 min. Kapa HiFi HotStart ReadyMix was incubated at 98 °C for 45 s before preparation of all PCR reactions (qPCR and final enrichment PCR), in order to activate the hot-start enzyme for successful nick translation at 72 °C in the first PCR step.

Final enrichment of the libraries was performed in a 50 µl reaction using 0.75 mM primers (custom Nextera primers as described for ATAC-seq) and 25 ml Kapa HiFi HotStart ReadyMix. Libraries were amplified for the number of cycles corresponding to the C_q value determined in the qPCR reaction. Enriched libraries were purified using SPRI beads (Beckman Coulter). To prepare input control samples, 3 µl of 50 mM $MgCl_2$ was added to 15 µl sonicated lysate (pool of 5 µl of endothelium, epithelium and fibroblast lysates from the same organ) to neutralize the EDTA in the SDS lysis buffer; 20 µl of tagmentation buffer and 1 µl transposase (Illumina) was added, and samples were incubated at 37 °C for 10 min; chromatin was purified with MinElute PCR purification kit (Qiagen), and 22.5 µl of the purified transposition reaction were combined with 25 µl of PCR master mix and 0.75 mM primers (custom Nextera). Control libraries were amplified as described above. Libraries were sequenced by the Biomedical Sequencing Facility at CeMM using the Illumina HiSeq 3000/4000 platform and the 50-bp single-end configuration. Epigenome mapping by ChIPmentation was done in two biologically independent experiments, with two exceptions: for endothelium from lymph node and fibroblasts from thymus, only one high-quality profile could be obtained. Sequencing statistics are provided in Supplementary Table 1.

Quantification of LCMV viral RNA

In the LCMV infection experiments, samples were collected at day 8 after infection and snap-frozen in liquid nitrogen. For detection of LCMV viral RNA, pieces of the organs were homogenized with a TissueLyser II (Qiagen), RNA was extracted using QIAzol lysis reagent (Qiagen), and reverse transcription was done using random primers and the First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The expression of the viral gene that encodes the LCMV nucleoprotein was measured with qPCR using TaqMan Fast Universal PCR Mastermix (Thermo Fisher Scientific) and previously published primers and probes[40]. The qPCR data were normalized against a reference comprising five established housekeeping genes (TaqMan Gene Expression Assays, Thermo Fisher Scientific): *Arf1* (Mm01946109_uH), *Rpl37* (Mm00782745_s1), *Rab1b* (Mm00504046_g1), *Ef1a* (Mm04259522_g1) and *Gapdh* (4352339E), to minimize biases introduced by potential tissue-specific expression of housekeeping genes[41,42]. Notably, this qPCR assay estimates infection levels based on LCMV viral RNA in bulk tissue, and it cannot account for organ-specific differences in the relative frequencies of cells that are susceptible to LCMV infection. The results should therefore be regarded as an indication (rather than as a precise quantitative measure) of the degree to which structural cells in different organs are directly and indirectly affected by the LCMV infections.

RNA-seq data processing and quality control

The RNA-seq data were processed and quality-controlled using established bioinformatics software and custom analysis scripts. Specific emphasis was put on ensuring high purity of the structural cell transcriptomes, while minimizing the risk and effect of potential contaminations (for example, as a result of cell-free RNA released by dying cells or potential impurities during FACS purification of the structural cell populations).

Raw reads were trimmed using trimmomatic (v.0.32) and aligned to the mouse reference genome (mm10) using HISAT2 (v.2.1.0). Gene expression was quantified by counting primary alignments to exons using the function 'summarizeOverlaps' from the GenomicAlignments package (v.1.6.3) in R (v.3.2.3). Gene annotations were based on the Ensembl GENCODE Basic set (genome-build GRCm38.p6).

To detect and remove potential biases arising from contaminating RNA of haematopoietic immune cells among the sorted structural cell populations, we screened all transcriptome profiles for gene expression signatures associated with eight types of haematopoietic immune cells. To that end, we derived gene signatures indicative of B cells, T cells, natural killer cells, natural killer T cells, macrophages, monocytes, dendritic cells and neutrophils from a recent mouse single-cell expression atlas[19]. In addition, a T cell receptor signature was defined by the genes *Cd3d*, *Cd3e*, *Cd3g*, *Lat*, *Lck*, *Vav1*, *Tbx21* and *Zap70*. For each of these gene signatures, aggregated expression values were calculated as follows. First, raw read counts were converted to counts per million (CPM) and $\log_2$-transformed using the function 'voom' from the limma package (v.3.26.9) in R. Second, voom-converted values were normalized to z-scores, followed by averaging across all genes in each gene signature. Samples with detectable contamination by haematopoietic immune cells were automatically discarded, and replacement samples were generated until three uncontaminated samples were available for each cell type in each organ.

As an additional precaution, we computationally corrected for residual contamination by RNA from haematopoietic immune cells. To that end, we regressed out the gene signatures of haematopoietic immune cells from the matrix of structural cell transcriptomes using the function 'removeBatchEffect' from the limma package in R. With this procedure, we generated corrected $\log_2$-transformed counts per million ($\log_2$(CPM)), which we used in all further analyses unless otherwise stated. Finally, for the comparisons between individual genes, we normalized the corrected expression values by gene length, thereby generating RPKM values.

Bioinformatic analysis of cell-type-specific and organ-specific gene expression

The processed and quality-controlled RNA-seq profiles of structural cells were analysed for characteristic differences in gene expression across cell types and organs, and enrichment analysis against public reference data was used to obtain an initial annotation of the identified gene signatures (Fig. 1d, Extended Data Fig. 3).

Cell-type-specific and organ-specific marker genes were identified using a two-step procedure. First, we performed pairwise differential expression analysis for each of the three type of structural cells, comparing them across organs. Second, based on the resulting pairwise comparisons, we identified for each organ those genes that were upregulated in each cell type compared to at least five other organs. Pairwise comparisons between organs were performed using the limma package in R, separately for each of the three structural cell types. Significantly differential genes were selected based on statistical significance (adjusted $P < 0.05$), average expression ($\log_2$(CPM) > -1) and sequencing coverage (median number of reads greater than 10 in the group with stronger signal). On the basis of these pairwise comparisons, we counted the total number of times each gene was upregulated in a specific organ compared to all other organs. Genes that were upregulated in comparison to five or more other organs were selected as marker genes of the corresponding organ.

The identified cell-type-specific and organ-specific marker genes were subjected to enrichment analysis with Enrichr (http://amp.pharm.mssm.edu/Enrichr/), using three Enrichr libraries: KEGG_2019_Mouse, GO_Biological_Process_2018 and Human_Gene_Atlas. The enrichment analysis provided an initial biological annotation of the identified marker genes (Fig. 1d) and validated their enrichment for previously reported cell-type-specific and organ-specific marker genes expressed in structural cells (Extended Data Fig. 3a). For further confirmation in an independent reference dataset[19], we normalized the expression values for each gene within each organ in our dataset, followed by averaging across genes for each set of marker genes from the reference dataset (Extended Data Fig. 3b). The resulting structural cell signatures thus reflect gene expression in one cell type compared to the other cell types in the same organ. Finally, we normalized these structural cell gene signatures across organs and cell types to *z*-scores, resulting in cell-type-specific and organ-specific gene signatures.

Bioinformatic inference of cell–cell interactions

To dissect the cell-type-specific and organ-specific crosstalk of structural cells with haematopoietic immune cells, we quantified the enrichment of known receptor–ligand pairs among the identified marker genes (Fig. 2a). Moreover, we assessed the expression of genes encoding receptors and ligands (Fig. 2b) and aggregated receptor and ligand gene signatures (Fig. 2c). Given the scale of our study, it was not possible to validate the inferred cell–cell interactions experimentally; however, the identification of many well-established interactions and of biologically plausible differences between cell types and organs provides support for these predictions.

Cell-type-specific and organ-specific marker genes of structural cells were obtained from the RNA-seq data as described in the previous section. Marker genes of haematopoietic immune cells were downloaded from supplementary table 4 of a recent mouse single-cell expression atlas[19]; from supplementary table 2 of the Tabula Muris paper[20], from data set_S02 (sheets 8, 12, 14 and 16) of a paper by the ImmGen consortium[43], and from supplementary table 1 of a large-scale characterization of tissue-resident macrophages[44]. To aggregate immune cell genes into a consensus set, the union of all identified marker genes was taken.

Receptor–ligand pairs were downloaded from CellPhoneDB[45] (as of 21 September 2018) and merged with additional receptor–ligand pairs extracted from supplementary data 2 of a recent paper[46], retaining only literature-supported pairs. Mouse gene identifiers were mapped to human gene identifiers based on the NCBI HomolGene mapping (build 68) using only the unique mappings. Interactions formed by receptor complexes from CellPhoneDB were transformed to pairwise interactions between individual receptors and ligands, by including an interaction between each member of a receptor complex and the respective ligand.

On the basis of these lists of marker genes and receptor–ligand pairs, we inferred potential cell–cell interactions for all pairs of one structural cell type and one haematopoietic immune cell type, quantifying the enrichment for known receptor–ligand pairs among all pairs of marker genes between the structural cell type and the haematopoietic cell type (Fig. 2a). First, we counted all pairs of marker genes for each pair of structural and immune cell types. Second, we calculated the fraction of these gene pairs that were annotated as receptor–ligand pairs. Third, we tested whether this fraction was greater than the fraction of annotated receptor–ligand pairs across all pairs of genes. Fisher's exact test was used to obtain *P* values and odds ratios, as implemented by the function 'fisher.test' in R. *P* values were adjusted for multiple testing using the Benjamini–Hochberg method, as implemented the function 'p.adjust' in R. Finally, significantly enriched pairs (adjusted $P < 0.05$) of structural and immune cell types were connected by edges to generate a graph of cell–cell interactions.

In addition to testing for significant enrichment (as described in the previous paragraph), we calculated aggregated receptor and ligand gene signatures (Fig. 2c) based on a manually curated list of immune-related receptors and ligands (Supplementary Table 4). These gene signatures were derived by normalizing expression values for each gene and averaging across all genes in a given set of biologically related receptors and ligands.

Bioinformatic analysis of the Tabula Muris dataset

To assess the expression patterns of immune genes in structural cells in an independent dataset (Extended Data Fig. 5), we obtained single-cell RNA-seq data and *t*-SNE projections from the Tabula Muris website https://tabula-muris.ds.czbiohub.org/ (21 December 2018) and cell counts from supplementary table 1 of the corresponding paper[20]. Cell types were assigned based on the provided cell ontology class, with stroma cells and basal cells of epidermis both labelled as fibroblasts. Gene expression counts were converted to transcripts per million (TPM) values and log-transformed. Receptor and ligand gene signatures in the Tabula Muris data were based on immune-related receptors and ligands (Supplementary Table 4) and normalized log(TPM) values for each gene, which were averaged across all genes in a given set of biologically related receptors and ligands.

ATAC-seq and ChIPmentation data processing and quality control

The ATAC-seq and H3K4me2 ChIPmentation data were processed using well-established bioinformatics pipelines, followed by quality control using the same approach as for the RNA-seq data (described above). The two data types were processed separately and subsequently integrated as described in the next section.

Raw reads were trimmed with trimmomatic (v.0.32) and aligned to the mouse reference genome (mm10) using bowtie2 (v.2.2.4). Primary alignments with mapping quality greater than 30 were retained. ATAC-seq peaks were called using MACS (v.2.7.6) on each individual sample. H3K4me2 ChIPmentation peaks were called using MACS against the input controls (which were obtained by pooling input control data from the three types of structural cells in equal amounts for each organ). Peaks were aggregated into a list of consensus peaks using the function *reduce* of the package GenomicRanges (version 1.22.4) in R. Consensus peak that overlapped with blacklisted genomic regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/) were discarded.

Quantitative measurements were obtained by counting reads within consensus peaks using the function 'summarizeOverlaps' from the GenomicAlignments (v.1.6.3) package in R. Samples with detectable contamination by haematopoietic immune cells were identified in the same way as for the RNA-seq data, using epigenomic signals in promoter peaks to calculate immune cell signatures. Contaminated samples were automatically removed and replaced by new samples. Finally, we computationally corrected for residual contamination in the retained samples by regressing out epigenomic signatures of haematopoietic immune cells from the matrix of signal intensity values across peaks, using the function 'removeBatchEffect' from the limma package in R.

Bioinformatic analysis of cell-type-specific and organ-specific transcription regulation

The processed and quality-controlled RNA-seq, ATAC-seq and H3K4me2 ChIPmentation data were analysed for differences in transcription regulation across cell types and organs (Fig. 3a–c). We integrated all three data types to derive sets of genomic regions with characteristic activity patterns (marker peaks), we performed motif enrichment analysis for marker peaks, and we inferred a gene-regulatory network by connecting enriched transcriptional regulators to their target genes. For the inference of gene-regulatory networks, we used DNA sequence motifs to connect transcriptional regulators to regulatory regions, and genomic proximity to connect regulatory regions to target genes. To make these coarse approximations more robust and interpretable, we used the ATAC-seq data to exclude motifs that are most likely inactive in the investigated cell types.

First, cell-type-specific and organ-specific marker peaks were identified in analogy to the RNA-seq-based identification of marker genes (described above), separately for the ATAC-seq and H3K4me2

ChIPmentation data. For each of the three cell types, we performed pairwise comparisons between organs using the limma package in R. Differential peaks were selected based on statistical significance (adjusted $P < 0.05$), signal intensity ($\log_2$(CPM) > −1) and sequencing coverage (median number of reads of at least 10 in the group with stronger signal). We then counted, separately for the two data types, the total number of times each peak showed increased signal in a specific organ compared with all other organs. Peaks that were upregulated compared to five or more other organs were selected as marker peaks of the corresponding organ.

Second, we derived a gene-linked list of marker peaks that were marked as promoter-associated peaks, enhancer-associated peaks and other peaks based on gene annotations as well as the RNA-seq and H3K4me2 ChIPmentation data (for enhancers). To that end, the ATAC-seq peaks were linked to all genes in which the transcription start site was located within 5 kb, using the function 'annotatePeakInBatch' of the package ChIPpeakAnno (v.3.4.6) package in R. They were annotated as promoter-associated if they were located within 200 bp from a transcription start site of the respective gene; or as enhancer-associated if they were located within 1–5 kb from a transcription start site, showed a correlation of ATAC-seq and RNA-seq signal greater than 0.3 and overlapped with an H3K4me2 ChIPmentation marker peak.

Third, we performed motif analysis on the gene-linked list of marker peaks to connect regulatory regions to transcriptional regulators. The HOMER software tool (v.4.9.1) was used to identify regulator binding motifs in the marker peaks and to determine regulator enrichment. This enrichment analysis was performed separately for all peaks, for promoter-associated peaks and for enhancer-associated peaks using the HOMER function 'findMotifsGenome' with a background set consistent of all gene-linked peaks. To avoid biases due to differences in peak size, all regions were standardized to 500 bp around the centre of the peak. Transcriptional regulators were labelled as significant based on statistical significance (adjusted $P < 0.001$) and the strength of the enrichment ($\log_2$(odds ratio) >1.5).

Fourth, we inferred a gene-regulatory network by connecting transcriptional regulators with their target genes. This analysis included only those regulators that were enriched in the motif analysis of gene-linked marker peaks (described in the previous paragraph). No regulators were excluded based on low or undetectable RNA expression levels, given that transcription factors can play a relevant regulatory role despite low expression levels. The gene-regulatory network was constructed as follows: (i) transcriptional regulators were linked to their target peaks using the HOMER function 'annotatePeaks'; (ii) the list of gene-linked marker peaks was used to map these peaks to their target genes; (iii) the final network was constructed based on the links between regulators and target peaks, and between peaks and target genes, using the package igraph (v.1.1.2) in R.

Fifth, we analysed the inferred gene-regulatory network by determining the network similarity between all pairs of transcriptional regulators using the function 'similarity' from the igraph package in R (with inverse log-weighted similarity). Similarities among transcriptional regulators were visualized by multidimensional scaling using the function 'cmdscale' in R, based on similarity measures normalized to a minimum of zero and maximum of one, and converted to distance measures by taking one minus the normalized similarity value.

Epigenetic potential for immune gene activation

To identify genes with unrealized epigenetic potential (that is, genes in which the chromatin state indicates much higher expression than observed under homeostatic conditions), we compared chromatin accessibility (ATAC-seq signal) of promoter regions with the expression levels of corresponding genes (RNA-seq signal) (Fig. 3d–g).

First, RNA-seq and ATAC-seq read counts were combined into a single table, converted to $\log_2$(CPM) values and quantile-normalized using the function 'voom' of the limma package in R. Second, we removed systematic differences between the two assays for each gene using the function 'removeBatchEffect' of the limma package in R. Third, we compared normalized RNA-seq and ATAC-seq signal intensities for each gene, organ and cell type using differential analysis from the limma package in R. This analysis used biological replicates for RNA-seq and ATAC-seq to statistically assess for a given combination of organ and cell type whether the ATAC-seq signal was disproportionately greater (compared to other organs and cell types) than the RNA-seq signal. On the basis of this analysis, genes were selected as significant based on the observed difference ($\log_2$-transformed fold change > 0.7), statistical significance (adjusted $P < 0.05$) and mean signal intensity (normalized $\log_2$(CPM) > 0). For ATAC-seq, we further required a minimum sequencing coverage (median number of reads > 5). In addition, we calculated the enrichment of genes with immune functions among genes with unrealized epigenetic potential. To that end, we selected the 200 genes with the greatest difference between ATAC-seq to RNA-seq signal in each organ and cell type, and we calculated their enrichment for our manually curated immune-related gene sets (Supplementary Table 8) using Fisher's exact test based on the function 'fisher.test' in R.

Bioinformatic analysis of the LCMV infection experiments

To evaluate the epigenetic potential for immune gene activation in structural cells, we compared the RNA-seq profiles of structural cells under homeostatic conditions with those collected on day 8 after infection with LCMV, and we assessed the predictive ability of the unrealized potential for the LCMV-induced changes in gene expression (Fig. 4). This analysis was based on the hypothesis that LCMV infection, which is widely regarded as a systemic infection model, will affect structural cells in most or all investigated organs, which may occur through direct effects (structural cells get infected) as well as indirect effects (structural cells respond to aspects of tissue-specific or systemic infection). Our gene expression analysis likely reflects a superposition of several such effects.

First, we identified the genes that were upregulated upon LCMV infection. To that end, we compared samples collected under homeostatic conditions with those collected after LCMV infection, separately for each cell type and organ, using the limma package in R. Differentially expressed genes were selected based on statistical significance (adjusted $P < 0.05$), average expression ($\log_2$(CPM) > 1) and sequencing coverage (median number of reads > 20 in the upregulated condition). Enrichment analysis was performed by comparing differential genes to our manually curated immune-related gene sets (Supplementary Table 8) using Fisher's exact test with the function 'fisher.test' in R.

Second, we evaluated whether genes with unrealized epigenetic potential under homeostatic conditions were preferentially upregulated in response to LCMV infection. For each cell type and organ, we split all genes into two groups based on their $\log_2$-transformed fold change comparing ATAC-seq signal to RNA-seq signal: Those with unrealized potential (ATAC-seq signal greater than RNA-seq signal) and those without unrealized potential (RNA-seq signal greater than ATAC-seq signal). We then assessed whether the LCMV-induced changes in gene expression ($\log_2$-transformed fold change comparing samples collected after LCMV infection to homeostatic conditions) was able to discriminate between the two groups defined by epigenetic potential. For visualization and quantitative comparison across cell types and organs, the data were plotted in diagrams that conceptually resemble receiver operating characteristic curves, and the predictive ability of the unrealized potential for gene activation afteer LCMV infection was quantified by area-under-the-curve values.

Third, the activated potential was calculated for each cell type and organ as the percentage of genes with unrealized potential under homeostatic conditions (ATAC-seq signal greater than RNA-seq signal, $P < 0.05$) that were among the significantly upregulated genes after LCMV infection (defined as described above).

Bioinformatic analysis of the cytokine treatment experiments

To dissect the in vivo effects of individual cytokines on structural cells (Fig. 5), we analysed the transcriptional response to the treatment of mice with six cytokines administrated as recombinant proteins, and we obtained RNA-seq profiles (processed and quality-controlled as described above) for analysis of gene expression.

First, we compared the gene expression of structural cells between cytokine-treated mice and mock-treated (PBS) controls using the limma package in R, separately for each cell type and organ. Differentially expressed genes were selected based on statistical significance (adjusted $P < 0.05$) and average expression ($\log_2$(CPM) >2).

Second, we compared the changes in gene expression upon cytokine treatment to the changes in gene expression upon LCMV infection. To that end, we correlated cytokine-related log-transformed fold changes to LCMV-related log-transformed fold changes across all genes, separately for each cell type, organ and cytokine.

Third, we tested whether genes upregulated after cytokine treatment were enriched for genes with unrealized epigenetic potential. To that end, we calculated the enrichment of genes with unrealized epigenetic potential (defined as described above) among genes upregulated by cytokine treatment (adjusted $P < 0.05$, log-transformed fold change > 0), separately for each organ, cell type and cytokine. Fisher's exact test was used to calculate statistical significance with the function 'fisher.test' in R.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw and processed sequencing data (RNA-seq, ATAC-seq and H3K4me2 ChIPmentation) are available from the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE134663. In addition, the dataset is provided as an online resource on a supplementary website (http://structural-immunity.computational-epigenetics.org), which includes links to raw and processed sequencing data, further analysis results and genome browser tracks for interactive visualization of the RNA-seq, ATAC-seq and ChIPmentation profiles. Source data are provided with this paper.

Code availability

The analysis source code underlying the final version of the paper is openly available from http://structural-immunity.computational-epigenetics.org.

32. Ahmed, R., Salmi, A., Butler, L. D., Chiller, J. M. & Oldstone, M. B. Selection of genetic variants of lymphocytic choriomeningitis virus in spleens of persistently infected mice. Role in suppression of cytotoxic T lymphocyte response and viral persistence. *J. Exp. Med*. **160**, 521–540 (1984).
33. Bergthaler, A. et al. Viral replicative capacity is the primary determinant of lymphocytic choriomeningitis virus persistence and immunosuppression. *Proc. Natl Acad. Sci. USA* **107**, 21641–21646 (2010).
34. ImmGen Consortium. Final ImmGen sorting SOP; http://www.immgen.org/Protocols/ImmGen%20Cell%20prep%20and%20sorting%20SOP.pdf (accessed 24 March 2020).
35. Krausgruber, T. et al. T-bet is a key modulator of IL-23-driven pathogenic CD4$^+$ T cell responses in the intestine. *Nat. Commun*. **7**, 11627 (2016).
36. Saluzzo, S. et al. First-breath-induced type 2 pathways shape the lung immune environment. *Cell Rep*. **18**, 1893–1905 (2017).
37. Lercher, A. et al. Type I interferon signaling disrupts the hepatic urea cycle and alters systemic metabolism to suppress T cell function. *Immunity* **51**, 1074–1087.e9 (2019).
38. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet*. **48**, 1193–1203 (2016).
39. Gustafsson, C., De Paepe, A., Schmidl, C. & Månsson, R. High-throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers. *BMC Genomics* **20**, 59 (2019).
40. Pinschewer, D. D. et al. Innate and adaptive immune control of genetically engineered live-attenuated arenavirus vaccine prototypes. *Int. Immunol*. **22**, 749–756 (2010).
41. Kouadjo, K. E., Nishida, Y., Cadrin-Girard, J. F., Yoshioka, M. & St-Amand, J. Housekeeping and tissue-specific genes in mouse tissues. *BMC Genomics* **8**, 127 (2007).
42. Li, B. et al. A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq. *Sci. Rep*. **7**, 4200 (2017).
43. Shay, T. et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl Acad. Sci. USA* **110**, 2946–2951 (2013).
44. Lavin, Y. et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326 (2014).
45. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
46. Ramilowski, J. A. et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun*. **6**, 7866 (2015).

Acknowledgements We thank the Core Facility Flow Cytometry of the Medical University of Vienna for cell sorting service; the Biomedical Sequencing Facility at CeMM for assistance with next-generation sequencing; S. Zahalka and S. Knapp for help and advice with the preparation of lung samples; S. Niggemeyer, J. Riede, S. Jungwirth and N. Fleischmann for animal caretaking; and all members of the Bock laboratory for their help and advice. This work was conducted in the context of two Austrian Science Fund (FWF) Special Research Programme grants (FWF SFB F6102; FWF SFB F7001). T.K. is supported by a Lise Meitner fellowship from the Austrian Science Fund (FWF M2403). N.F. is supported by a fellowship from the European Molecular Biology Organization (EMBO ALTF 241-2017). A.L. is supported by a DOC fellowship of the Austrian Academy of Sciences. A.B. is supported by an ERC Starting Grant (European Union's Horizon 2020 research and innovation programme, grant agreement no. 677006). C.B. is supported by a New Frontiers Group award of the Austrian Academy of Sciences and by an ERC Starting Grant (European Union's Horizon 2020 research and innovation programme, grant agreement no. 679146).

Author contributions T.K. designed the project, performed experiments, analysed data and cowrote the manuscript. N.F. designed and performed the bioinformatic analysis and cowrote the manuscript. V.F.-G., M.S., L.C.S., A.N. and C.S. contributed to sample collection and sequencing library preparation. A.L. contributed to the experimental design and performed in vivo experiments (LCMV infections; cytokine treatments). A.F.R. contributed bioinformatic software. A.B. contributed to the experimental design and supervised the in vivo experiments. C.B. supervised the project and cowrote the manuscript. All authors read, contributed to, and approved the final manuscript.

Competing interests The authors declare no competing interests.

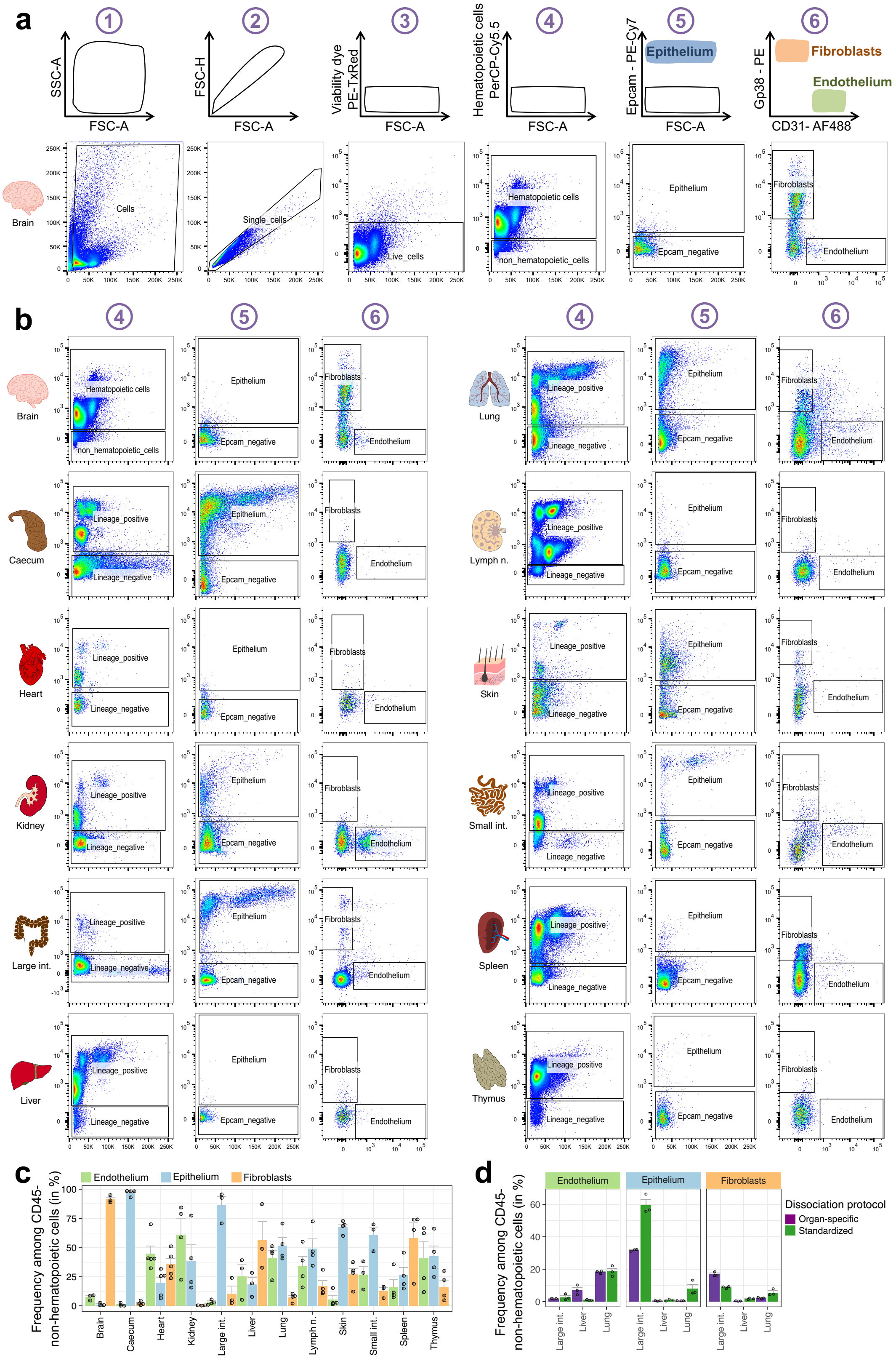
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2424-4.

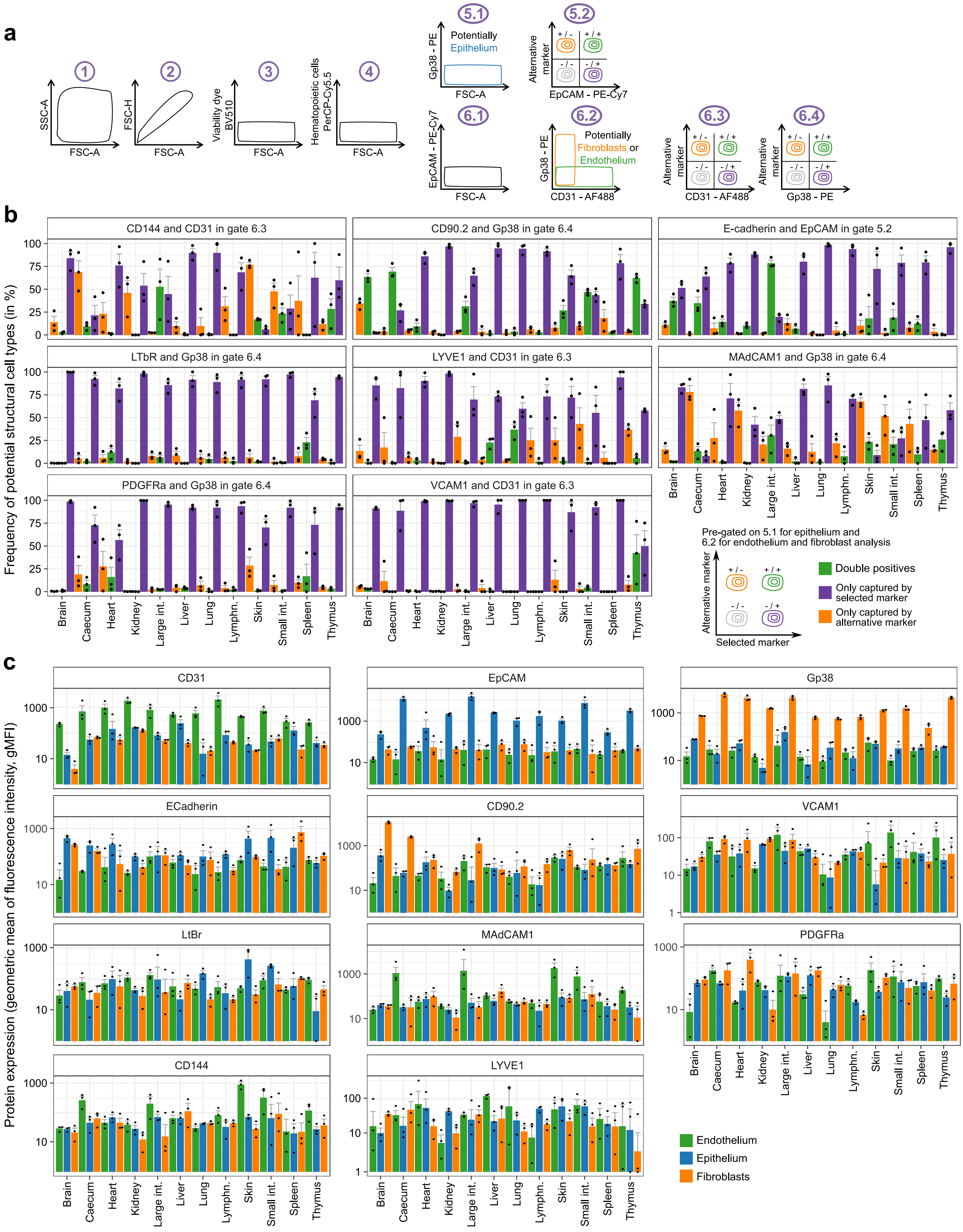
Correspondence and requests for materials should be addressed to C.B.

Peer review information *Nature* thanks Ari M. Melnick and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

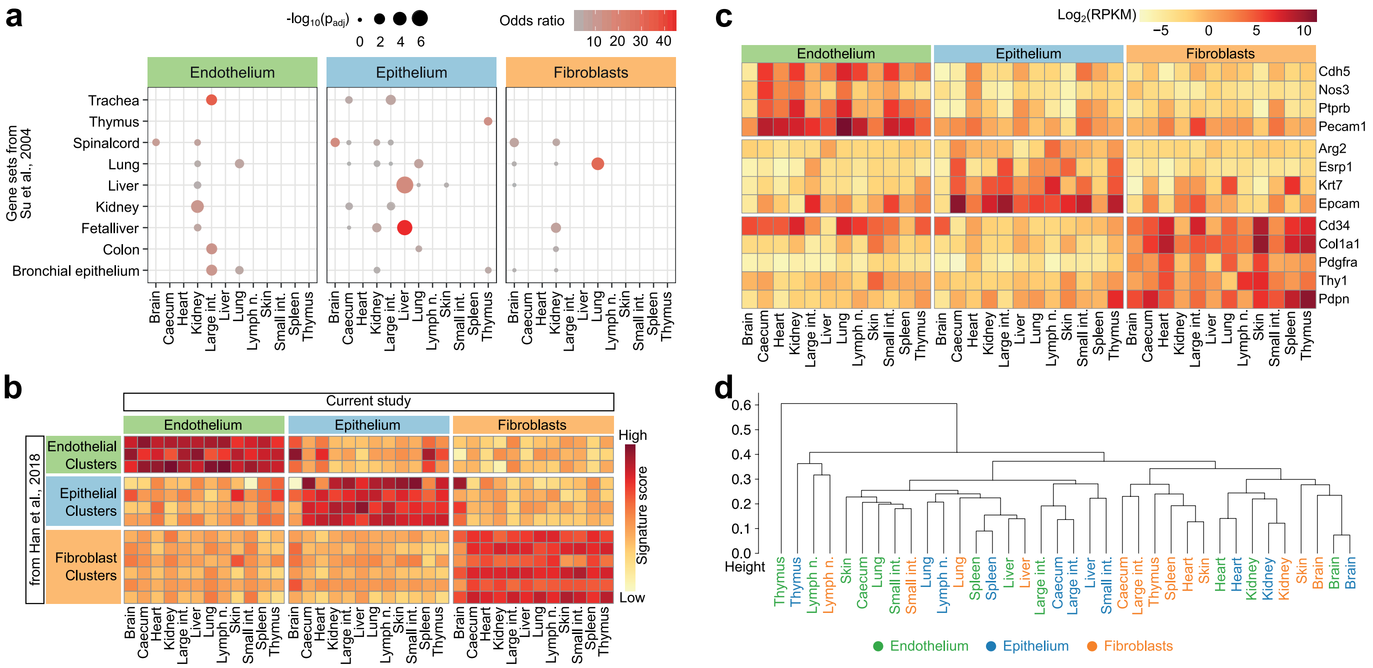
Reprints and permissions information is available at http://www.nature.com/reprints.



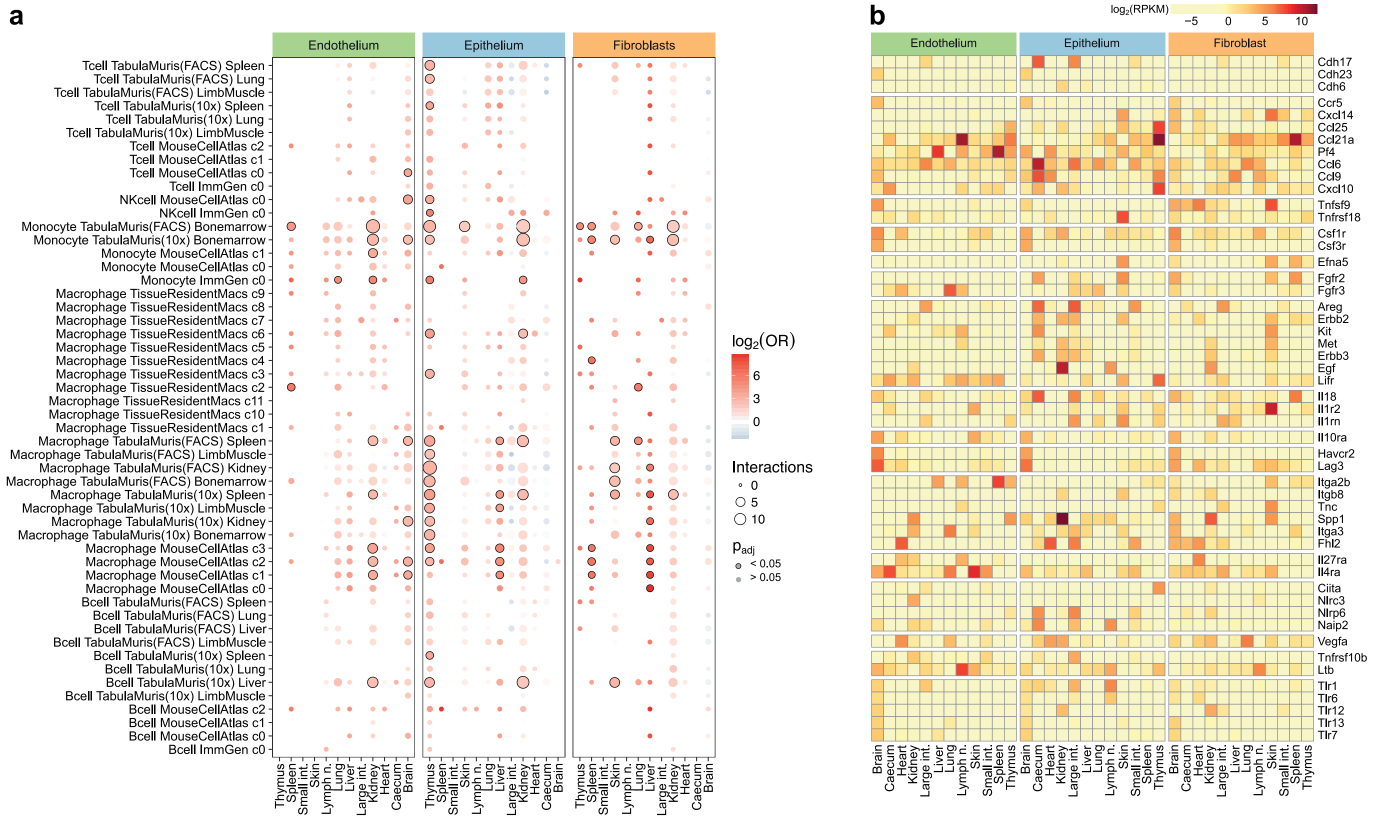
Extended Data Fig. 1 | Standardized identification and purification of structural cells across 12 organs. a, Cell-type identification and cell-sorting scheme (top row) with representative flow cytometry plots (selected from $n = 4$ independent biological replicates) in one representative organ (brain) under homeostatic conditions (bottom row). **b**, Representative plots (selected from $n = 4$ independent biological replicates) for gating steps 4–6 of the standardized cell-type identification and cell-sorting scheme (**a**) across the 12 organs under homeostatic conditions. **c**, Relative frequencies of structural cell types among non-haematopoietic (CD45$^-$) cells across 12 organs, for cell suspensions obtained by standardized organ dissociation. **d**, Relative frequencies of structural cell types among non-haematopoietic (CD45$^-$) cells across three organs, for cell suspensions obtained by either standardized organ dissociation or organ-specific dissociation protocols. Shown are mean and s.e.m. values. Sample size: $n = 4$ (**c**) and $n = 3$ (**d**) independent biological replicates.



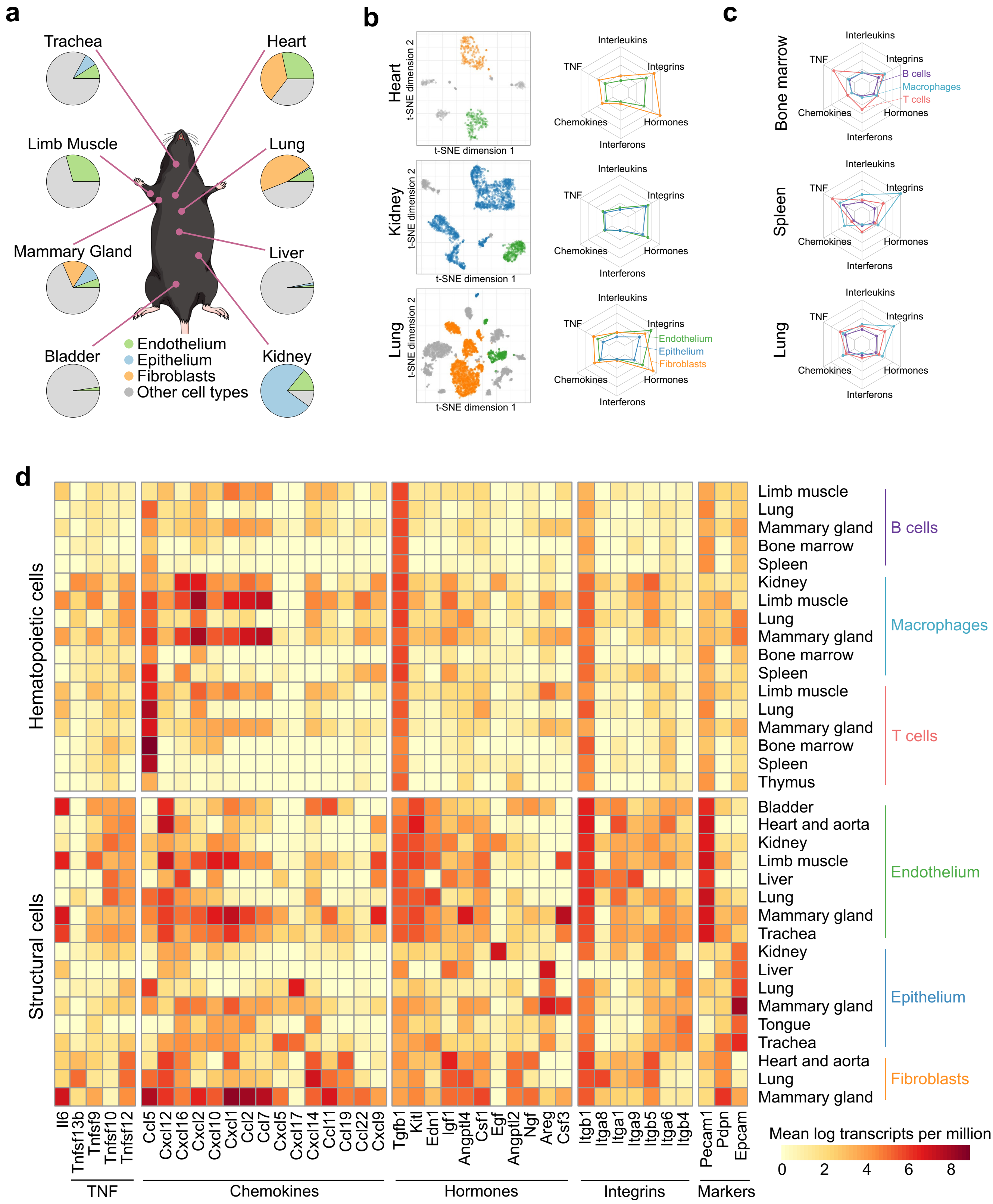
Extended Data Fig. 2 | Surface marker profiling of structural cells under homeostatic conditions. a, Gating strategy for the flow cytometry-based validation of the structural cell sorting scheme. Identification of structural cells starts with gating for intact cells (1), single cells (2), live cells (3) and non-haematopoietic cells (4). From the resulting non-haematopoietic (CD45$^-$) cell population, potential epithelial cells (5.1) are gated for epithelial cell markers (5.2). Similarly, potential endothelial cells and fibroblasts (6.1, 6.2) are gated for endothelial cell markers (6.3) and fibroblast markers (6.4). **b**, Relative frequencies of potential structural cell types based on gates 5.2, 6.3 and 6.4 (from **a**), comparing the selected markers with alternative markers. **c**, Expression of the selected surface markers of structural cell types (top row) and potential alternative markers for cells gated as in Extended Data Fig. 1a. Shown are mean and s.e.m. values. Sample size (all panels): *n* = 3 independent biological replicates.



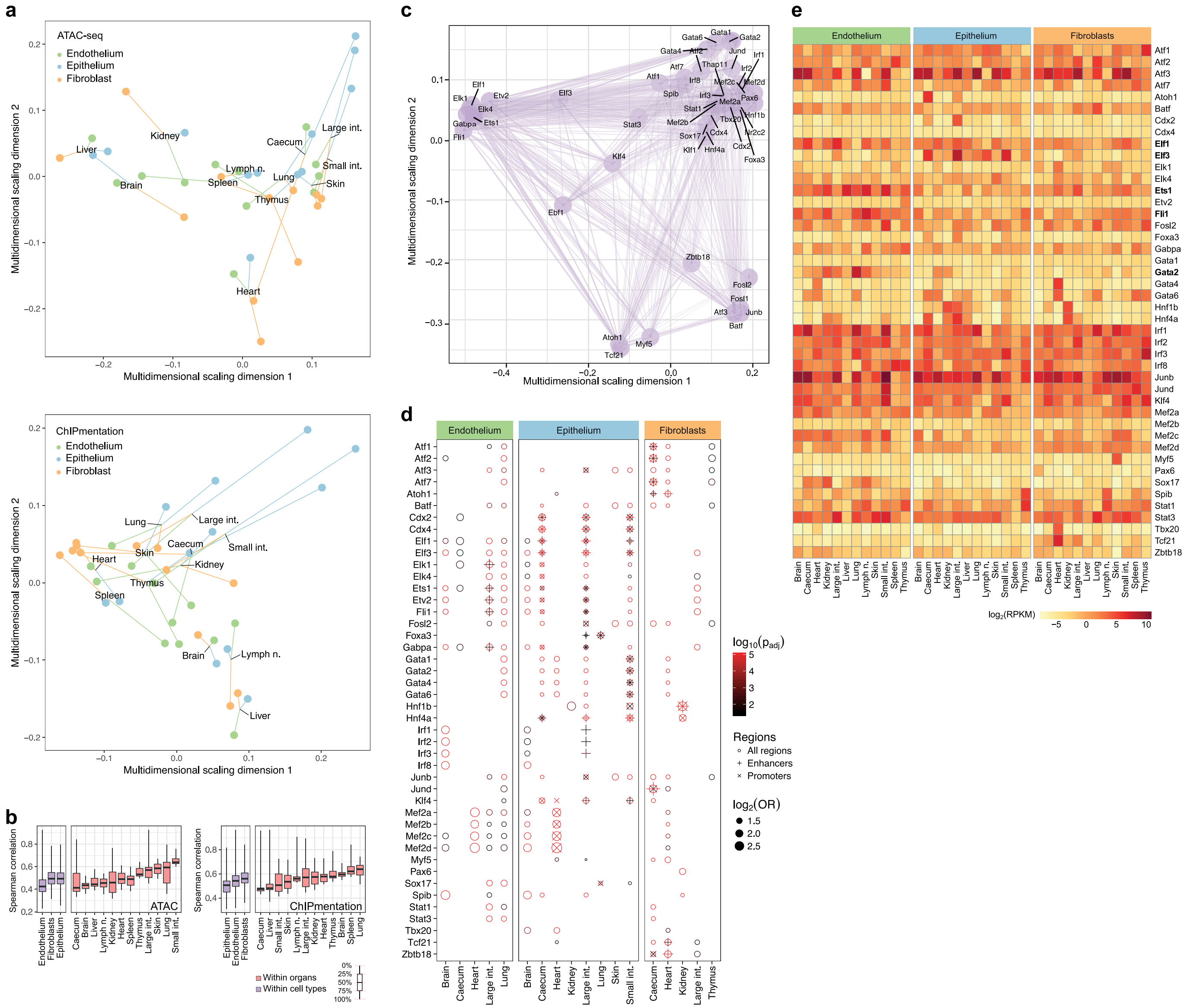
Extended Data Fig. 3 | Comparison of the structural cell transcriptomes to published reference data. a, Overlap of the identified cell-type-specific and organ-specific marker genes (derived from the RNA-seq experiments in the current study) with tissue-specific gene sets from a microarray-based expression atlas (two-sided Fisher's exact test with multiple-testing correction). **b**, Gene expression across cell types and organs (from the current study) aggregated across marker genes of structural cell clusters in a single-cell RNA-seq atlas of the mouse[19]. **c**, Gene expression across cell types and organs (from the current study) plotted for a manually curated list of commonly used markers of structural cells. **d**, Hierarchical clustering of structural cells across cell types and organs based on the transcriptome profiles from the current study. Sample size (all panels): *n* = 3 independent biological replicates.



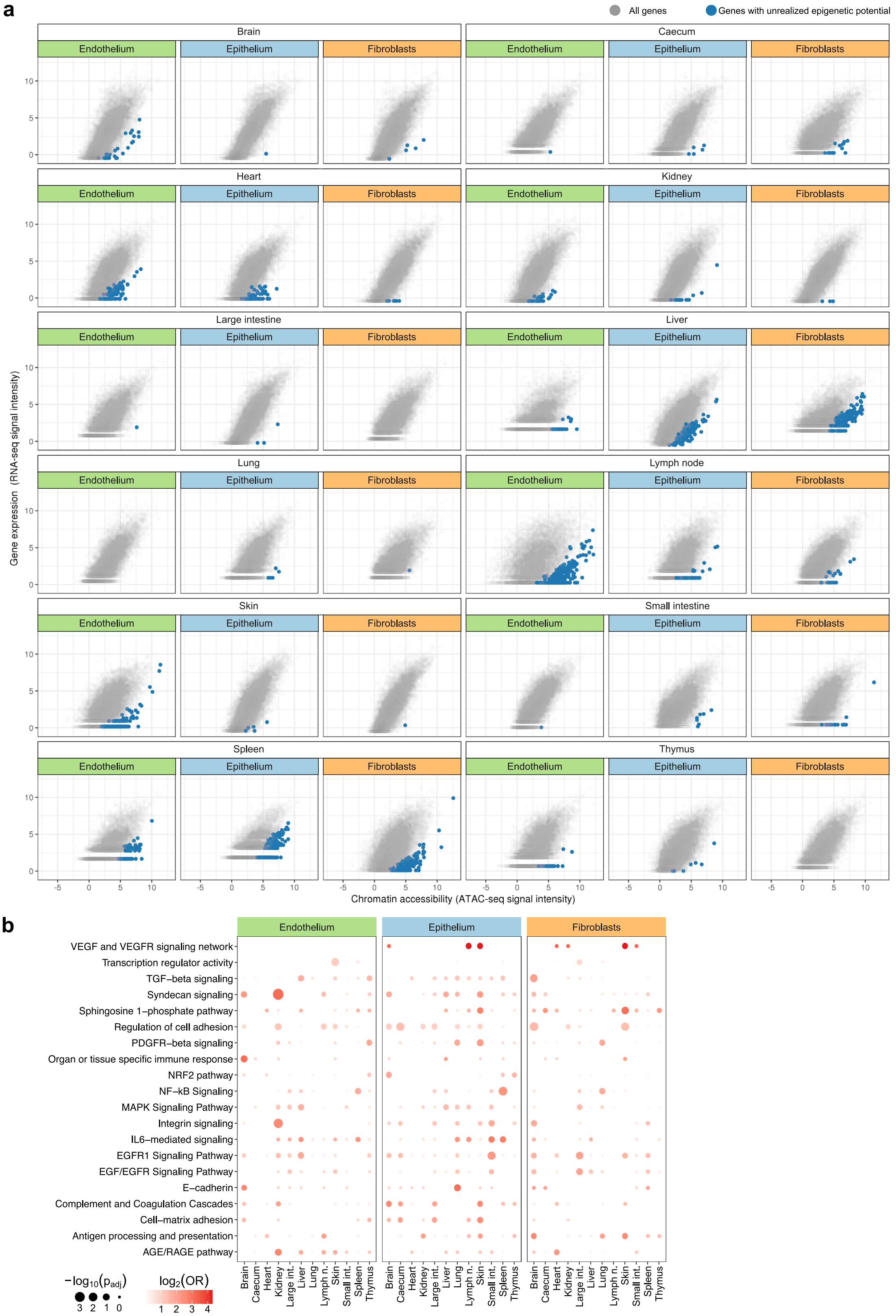
Extended Data Fig. 4 | Inference of cell–cell interactions across cell types and organs. a, Enrichment analysis for potential cell–cell interactions between structural cells and haematopoietic immune cells, based on gene expression of known receptor–ligand pairs (two-sided Fisher's exact test with multiple-testing correction). For each combination of one structural cell type and one haematopoietic immune cell type, the analysis assesses whether all pairs of marker genes between the two cell types are enriched for annotated receptor–ligand pairs. **b**, Differently expressed genes across cell types and organs, based on a manually curated list of receptors and ligands (Supplementary Table 4). Sample size (all panels): *n* = 3 independent biological replicates.



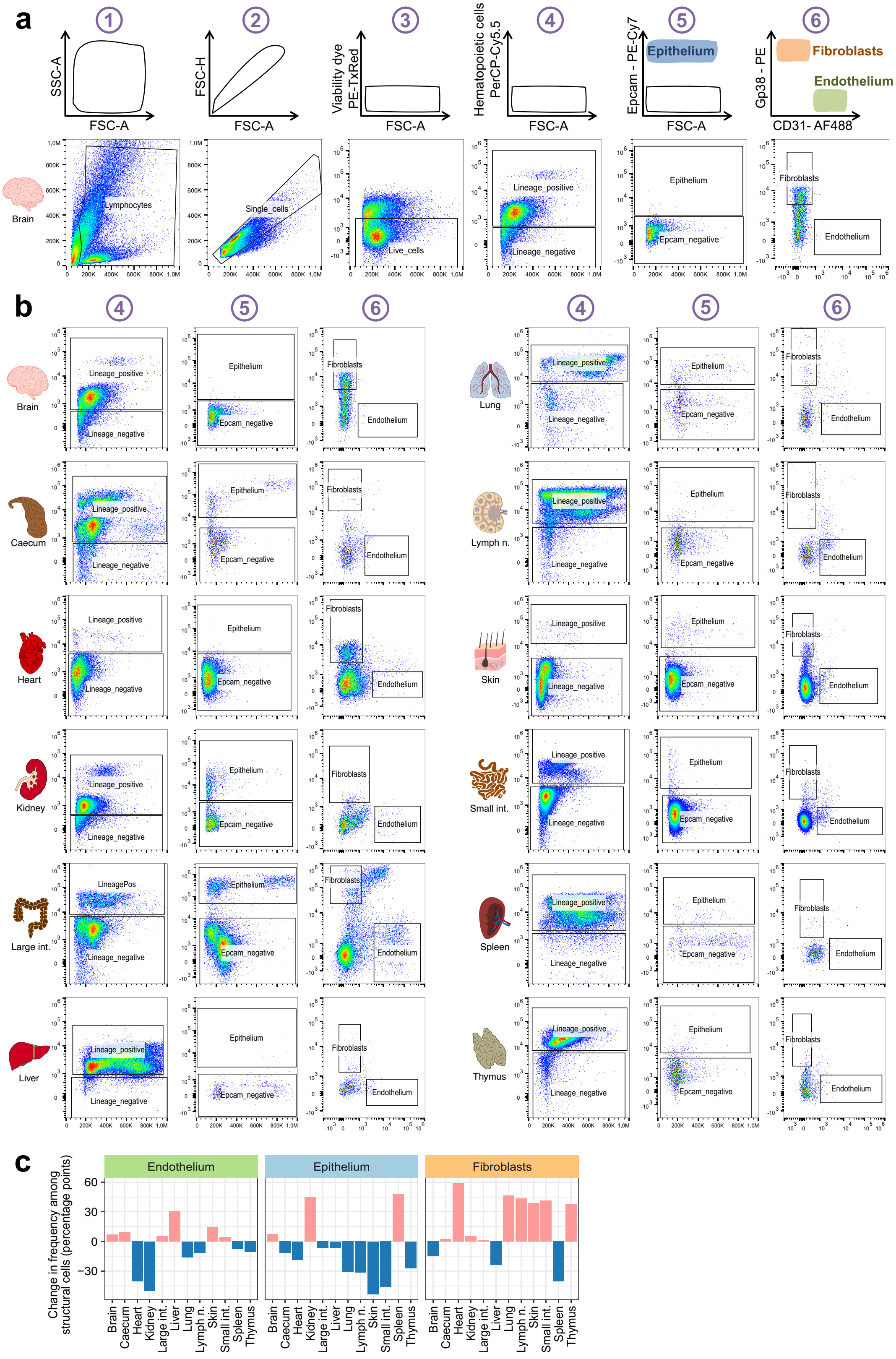
Extended Data Fig. 5 | Analysis of immune gene expression among structural cells in an independent dataset. **a**, Relative frequencies of single-cell transcriptomes classified as endothelium, epithelium and fibroblasts in selected organs according to the Tabula Muris dataset[20]. **b**, Expression of immune gene signatures in structural cells according to the Tabula Muris dataset, jointly normalized across all plots (for comparability). **c**, Expression of immune gene signatures in haematopoietic immune cells according to the Tabula Muris dataset, normalized as in **b**. **d**, Expression of selected immune genes in structural cells and in haematopoietic immune cells according to the Tabula Muris dataset. Sample size: *n* = 7 (all panels) independent biological replicates, comprising 4 male and 3 female mice.



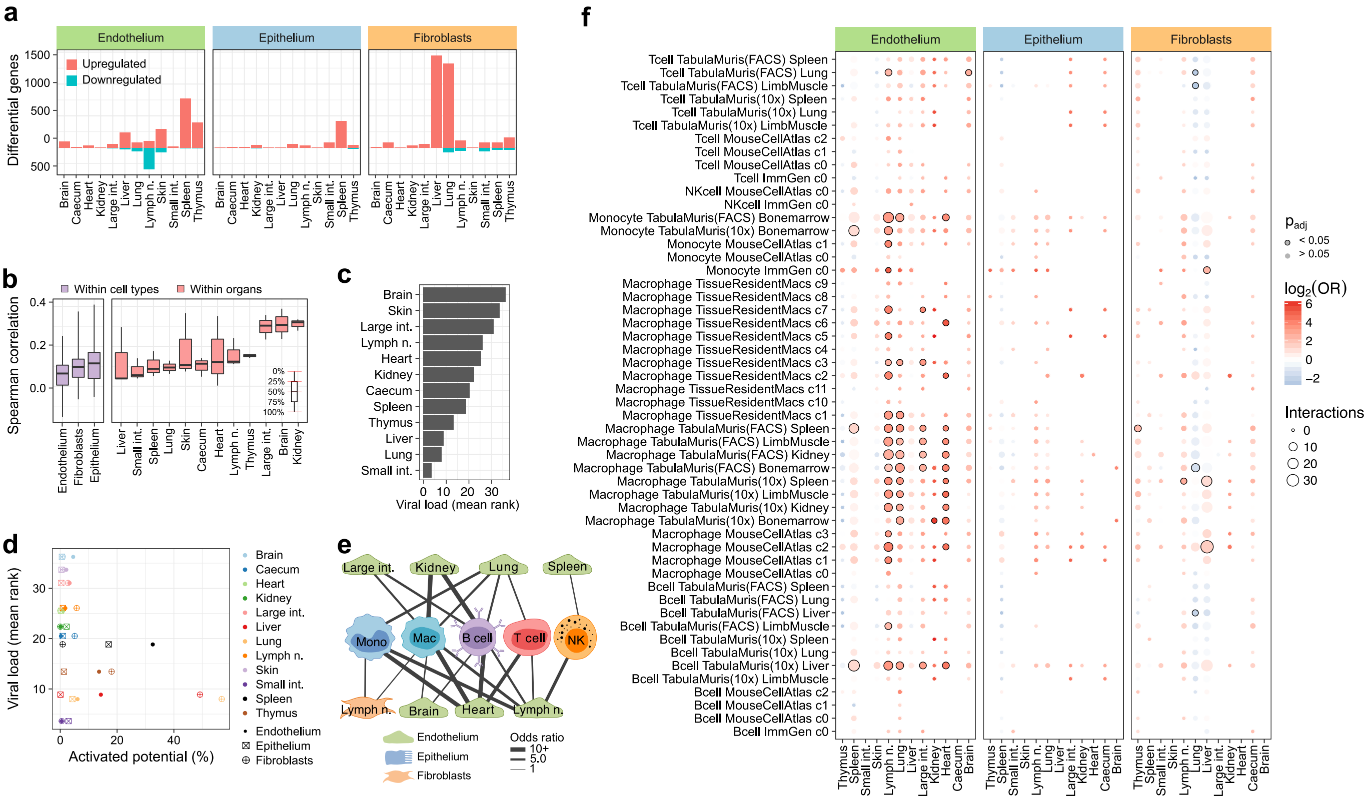
Extended Data Fig. 6 | Analysis of transcription regulation in structural cells. a, Multidimensional scaling analysis of the similarity of chromatin profiles across cell types, organs and replicates based on ATAC-seq (top) and H3K4me2 ChIPmentation (bottom). **b**, Correlation of chromatin profiles across cell types and organs for ATAC-seq (left) and H3K4me2 ChIPmentation (right). **c**, Transcriptional regulators of the inferred gene-regulatory network for structural cells, arranged by similarity using multidimensional scaling. **d**, Motif enrichment for transcriptional regulators among differential chromatin peaks, shown separately for each regulator (one-sided hypergeometric test with multiple-testing correction). **e**, Gene expression of the transcriptional regulators across cell types and organs (genes discussed in the text are in bold). Sample size (all panels): $n = 2$ independent biological replicates.



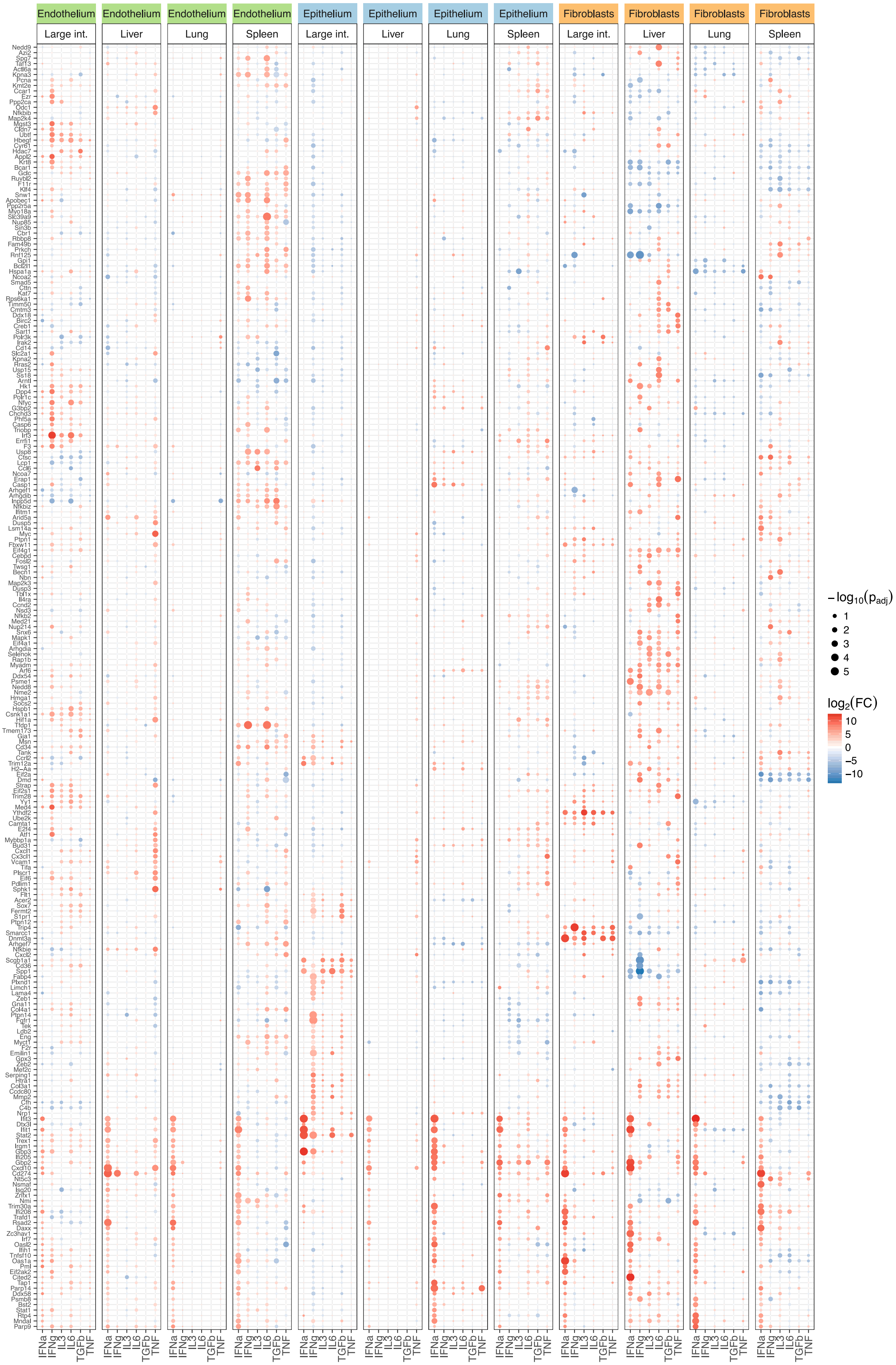
Extended Data Fig. 7 | Detection and analysis of genes with unrealized epigenetic potential. **a**, Scatterplot showing the correlation between chromatin accessibility in promoter regions and the corresponding gene expression levels in structural cells across cell types and organs. Genes with significant unrealized epigenetic potential (calculated as the difference between normalized ATAC-seq and RNA-seq signals) are highlighted in blue. **b**, Enrichment of immune-related gene sets among the genes with unrealized epigenetic potential (two-sided Fisher's exact test with multiple-testing correction). Sample size (all panels): *n* = 2 independent biological replicates.



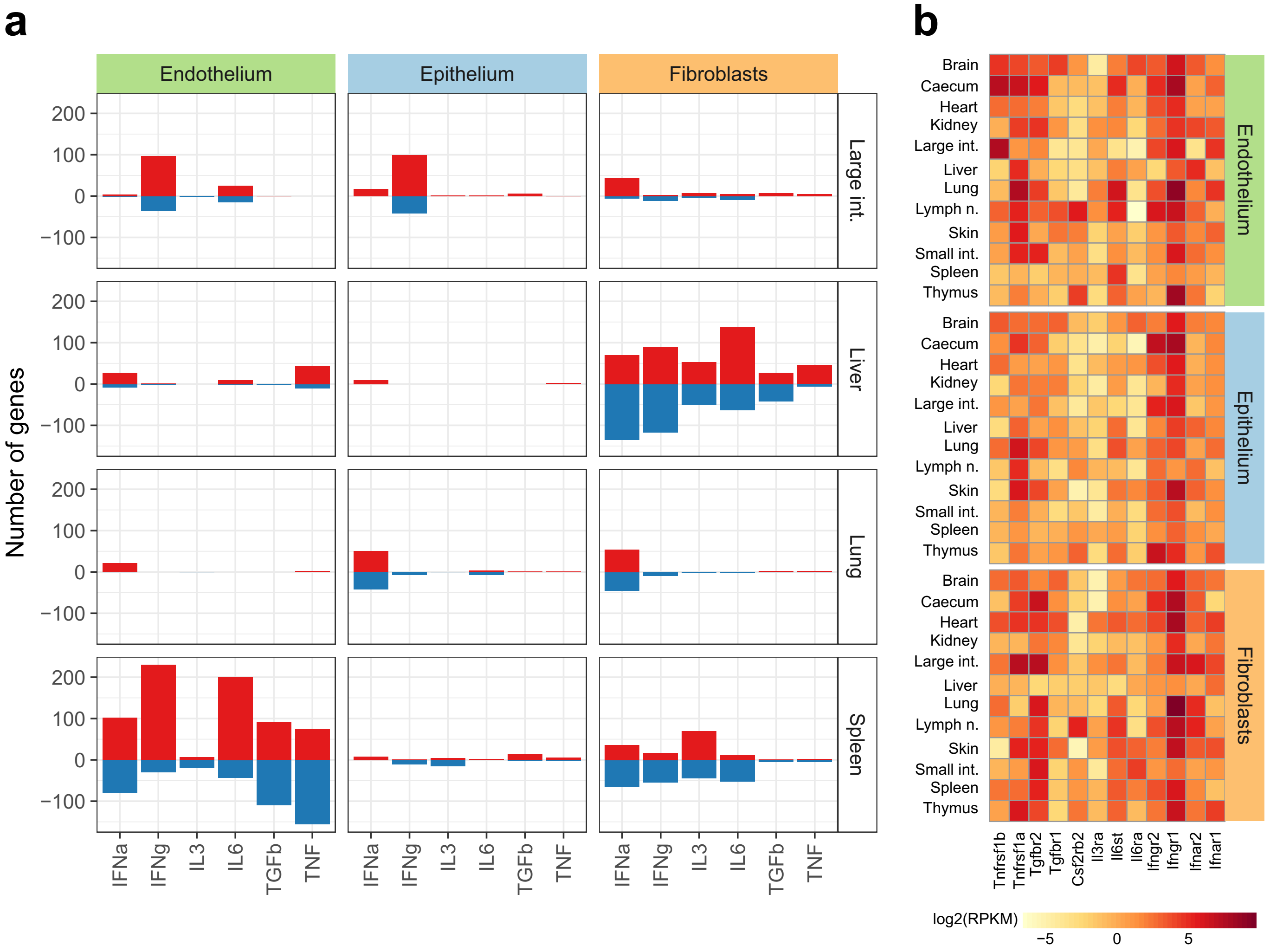
Extended Data Fig. 8 | Standardized identification and purification of structural cells after LCMV infection. a, Cell-type identification and cell-sorting scheme (top row) with representative flow cytometry plots (selected from $n = 3$ independent biological replicates) in one representative organ (brain) after LCMV infection (bottom row). **b**, Representative plots (selected from $n = 3$ independent biological replicates) for gating steps 4–6 of the standardized cell-type identification and cell-sorting scheme (**a**) across the 12 organs after LCMV infection. **c**, Change in the relative frequency of structural cells after LCMV infection. Sample size (all panels): $n = 3$ independent biological replicates.



Extended Data Fig. 9 | Analysis of differential gene expression in response to LCMV infection. a, Number of differentially expressed genes in structural cells after LCMV infection (this includes not only immune genes but also genes associated with the substantial organ-specific tissue damage and other direct and indirect effects of LCMV infection). **b**, Correlation of the observed changes in gene expression after LCMV infection across cell types and organs. **c**, Organ-specific viral load at day 8 of LCMV infection, measured by qPCR in whole-tissue samples collected from each organ (without FACS purification of individual cell types). Five reference genes were used for normalization and results were ranked across organs, to make the analysis robust towards tissue-specific differences in the expression of these housekeeping genes. However, the experimental results do not support an absolute quantification of viral load in each organ nor do they account for differences in the relative frequencies of cells that are susceptible to LCMV infection across organs. **d**, Scatterplot illustrating the low correlation between the activated epigenetic potential and the measured viral load across cell types and organs. **e**, **f**, Network analysis (**e**) and enrichment analysis (**f**) of potential cell–cell interactions between structural cells and haematopoietic immune cells, inferred from gene expression of known receptor–ligand pairs after LCMV infection (two-sided Fisher's exact test with multiple-testing correction). For each combination of one structural cell type and one haematopoietic immune cell type, the analysis assesses whether all pairs of marker genes between the two cell types are enriched for annotated receptor–ligand pairs. Sample size (all panels): *n* = 3 independent biological replicates.



Extended Data Fig. 10 | Visualization of differential gene expression in response to in vivo cytokine treatments. The heat map displays changes in the expression of genes associated with immune functions, plotted across cell types, organs and cytokines (two-sided linear model with multiple-testing correction). Sample size: $n = 3$ independent biological replicates.



Extended Data Fig. 11 | Analysis of differential gene expression in response to in vivo cytokine treatments. a, Number of differentially expressed genes in response to the individual cytokine treatments. **b**, Gene expression for the known receptors involved in the response to the individual cytokine treatments, plotted across cell types and organs under homeostatic conditions. Sample size (all panels): *n* = 3 independent biological replicates.

nature research

Corresponding author(s): Christoph Bock

Last updated by author(s): Apr 21, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Flow cytometry data were collected with DB FACSDiva (version 8.0.1) and sort-purification of cells was performed with Summit (version 6.3.1.16945; MoFlo Astrios cell sorter) or Sony Cell Sorter (version 2.1.5, SH800 cell sorter) software. For determining optimal enrichment cycles during ATAC-seq and ChIPmentation library preparation, qPCR data were collected using Bio-Rad CFX Manager (version 3.1).
Data analysis	The following software packages were used for data analysis: Flow cytometry data were analyzed with FlowJo (Version 10.5.3 Tree Star) software. Raw sequencing data were processed using trimmomatic (version 0.32), HISAT2 (version 2.1.0), bowtie2 (version 2.2.4), and MACS (version 2.7.6). Integrated data analysis was performed in R (version 3.2.3), based on the packages ChIPpeakAnno (3.4.6), data.table (1.11.4), DiffBind (1.16.3), doMC (1.3.5), fmsb (0.6.3), foreach (1.4.4), gdata (2.18.0), GenomicRanges (1.22.4), ggplot2 (2.2.1), gplots (3.0.1), grid (3.2.3), gridExtra (2.3), igraph (1.1.2), limma (3.26.9), LOLA (1.0.0), Matrix (1.2.14), pheatmap (1.0.10), project.init (0.0.1), RColorBrewer (1.1.2), readxl (1.3.1), ROCR (1.0.7), Rsamtools (1.22.0), scales (0.5.0), and WriteXLS (5.0.0). Inkscape (Version 0.92) was used for data visualization and figure preparation. The analysis source code underlying the final version of the paper is available from the Supplementary Website (http://structural-immunity.computational-epigenetics.org).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw and processed sequencing data (RNA-seq, ATAC-seq, H3K4me2 ChIPmentation) are available from the NCBI Gene Expression Omnibus (GEO) repository (accession number: GSE134663). In addition, the data are provided as an online resource on a supplementary website (http://structural-immunity.computational-epigenetics.org), which includes links to the raw and processed data, further analysis results, and genome browser tracks for interactive visualization of the RNA-seq, ATAC-seq, and ChIPmentation profiles.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was based on established standards in the field, including guidelines and SOPs from the Immgen project.
Data exclusions	To avoid potential biases due to contaminating RNA from hematopoietic immune cells (which may be released from dying cells or as the result of potential impurities during FACS purification of the structural cell populations), we screened all transcriptome profiles for gene expression signatures associated with hematopoietic immune cells. Samples with detectable contamination were automatically discarded, and replacement samples were generated until three uncontaminated samples were available for each cell type in each organ. These exclusion criteria were not pre-established; rather, they were devised based on the data for the first sample batches and consistently applied thereafter.
Replication	Transcriptome profiling was done in three biologically independent experiments. Chromatin accessibility mapping by ATAC-seq was done in two biologically independent experiments. Epigenome mapping by ChIPmentation was done in two biologically independent experiments, with two exceptions: for endothelium from lymph node and fibroblasts from thymus, only one high-quality sample could be obtained.
Randomization	Processing of samples from the various organs and cell types did not follow any particular order. The dataset was monitored for evidence of batch effects, but these analyses did not identify any technical covariates that should be controlled for using statistical methods.
Blinding	Because there were no pre-defined sample groups during data collection, blinding was not applicable for this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☐	☒	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	CD16/CD32 blocking antibody (clone 93, BioLegend Cat# 101320, Lot# B264827, dilution 1:200) CD31 (clone MEC13.3, AF488, BioLegend Cat# 102514, Lot# B282351, dilution 1:100) CD45 (clone 30-F11, PerCP-Cy5.5, BioLegend Cat# 103132, Lot# B295198, dilution 1:300) CD90.2 (clone 30-H12, AF700, BioLegend Cat# 105319, Lot# B260794, dilution 1:100) CD106 (clone 429, AF647, BioLegend Cat# 105711, Lot# B269055, dilution 1:100) CD140a (clone APA5, BV605, BD Bioscience Cat# 704380, Lot# 9284670, dilution 1:100) CD144 (clone BV13, BV421, BioLegend Cat# 138013, Lot# B292996, dilution 1:100) CD324 (clone DECMA-1, APC-Cy7, BioLegend Cat# 147314, Lot# B274948, dilution 1:150) Epcam (CD326) (clone G8.8, PE-Cy7, BioLegend Cat# 118216, Lot# B246071, dilution 1:150) MAdCAM1 (clone MECA-367, BV421, BD Bioscience Cat# 742812, Lot# 9284682, dilution 1:100) Lymphotoxin beta receptor (clone 5G11, APC, BioLegend Cat# 134408, Lot# B283402, dilution 1:100) LYVE1 (clone ALY7, eFlour 660, ThermoFisher Cat# 50-0443-82, Lot# 2107832, dilution 1:100) Podoplanin (gp38) (clone 8.1.1, PE, BioLegend Cat# 127408, Lot# B296330, dilution 1:150) Ter119 (clone TER119, PerCP-Cy5.5, BioLegend Cat# 116228, Lot# B291958, dilution 1:200) Dead cells were stained by adding Zombie Red Fixable Viability Dye (BioLegend Cat# 423110, Lot# B225035, dilution 1:1000) or Zombie Aqua Fixable Viability Dye (BioLegend Cat# 423102, Lot# B274475, dilution 1:1000)

Validation

CD16/CD32 blocking antibody: Quality tested by flow cytometric analysis according to vendor's website.
CD31 (clone MEC13.3): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse splenocytes) shown on vendor's website.
CD45 (clone 30-F11): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse splenocytes) shown on vendor's website.
CD90.2 (clone 30-H12): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse thymocytes) shown on vendor's website.
CD106 (clone 429): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse bone marrow myeloid cells) shown on vendor's website.
CD140a (clone APA5): Quality tested by flow cytometric analysis of antibody surface-stained cells according to vendor's website.
CD144 (clone BV13): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse endothelial cells bEnd.3) shown on vendor's website.
CD324 (clone DECMA-1): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of MDCK epithelial cell line) shown on vendor's website.
Epcam (clone G8.8): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse thymic epithelial stromal cell line TE-71) shown on vendor's website.
MAdCAM1 (clone MECA-367): Quality tested by flow cytometric analysis of antibody surface-stained cells according to vendor's website.
Lymphotoxin beta receptor (clone 5G11): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse bone marrow cells) shown on vendor's website.
LYVE1 (clone ALY7): Quality tested by flow cytometric analysis of mouse tissue, flow cytometry plots shown on vendor's website.
Podoplanin (clone 8.1.1): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse thymic epithelial stromal cell line TE-71) shown on vendor's website.
Ter119 (clone TER119): Quality tested by flow cytometric analysis of antibody surface-stained cells, flow cytometry plot (staining of mouse bone marrow cells) shown on vendor's website.
Zombie Red Fixable Viability Dye: Quality tested by flow cytometric analysis, flow cytometry (staining of one day old mouse splenocytes) shown on vendor's website.
Zombie Aqua Fixable Viability Dye: Quality tested by flow cytometric analysis, flow cytometry (staining of one day old mouse splenocytes) shown on vendor's website.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

C57BL/6J mice were bred and maintained under specific pathogen free conditions at the Institute of Molecular Biotechnology (IMBA) of the Austrian Academy of Sciences in Vienna (Austria). Experiments were performed under specific pathogen free conditions at the Anna Spiegel Research Building, Medical University of Vienna (Austria). Age-matched male mice (8 to 13 weeks old) were used for the characterization under homeostatic conditions, for the viral infection model, and for in vivo cytokine treatments.

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected in the field.

Ethics oversight

All mouse experiments were done in compliance with the respective animal experiment licenses (BMWFW-66.009/0199-WF/V/3v/2015 and BMWFW-66.009/0361-WF/V/3b/2017) approved by the institutional ethical committees and the institutional guidelines at the Department for Biomedical Research of the Medical University of Vienna.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134663

Files in database submission

Raw read counts; normalized log2-transformed counts-per-million

Genome browser session
(e.g. UCSC)

http://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&hubClear=http://medical-epigenomics.org/papers/krausgruber2019/hub.txt&hgt.labelWidth=30&pix=1300&textSize=12&position=chr16:91,542,434-91,570,261

Methodology

Replicates

Two biological replicates

Sequencing depth

12 samples per HiSeq 3000/4000 lane were sequenced, with ~360 million reads per lane, thus aiming for 30 million reads per samples. 16 million to 50 million reads per sample were obtained. Sequencing was done with 50 bp single-end reads.

Antibodies	H3K4me2 (clone AW30, Merck Cat#04-790)
Peak calling parameters	Peaks were called using MACS2 (version 2.7.6) function callpeak, comparing samples to inputs controls. The input controls were obtained by pooling sort-purified cells (endothelium, epithelium, fibroblast) in equal amounts for each organ.
Data quality	Data quality including peak height and background noise was assessed by visual inspection using the UCSC Genome Browser.
Software	Peak lists were aggregated to a consensus peak list using the function reduce of the package GenomicRanges (version 1.22.4) in R. Consensus peaks overlapping with blacklisted regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/) were discarded. Quantitative measurements were obtained by counting reads within consensus peaks using the function summarizeOverlaps from the GenomicAlignments (version 1.6.3) package in R.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Standardized sample collection and organ dissociation
Different surface markers and sorting schemes were previously used to purify endothelium, epithelium, and fibroblasts in individual organs, while our systematic comparison required standardized cell purification to be informative. We therefore tested multiple surface markers across organs and optimized a sorting scheme that produced consistent results across all 12 investigated organs, while excluding cell types that were detectable only in one or a few tissues (most notably lymphatic endothelial cells). The experimental workflow followed the recommendations of the Immunological Genome Project regarding sample collection schedule, antibody staining, and sample pooling. At least three same-sex littermate mice were pooled for each biological replicates for homeostatic conditions and for LCMV infection, respectively. For the in vivo cytokine treatments, we used individual mice as biological replicates to reduce the total number of mice. Standardized organ harvesting and dissociation protocols were established to obtain single-cell suspension for subsequent cell purification by fluorescence activated cell sorting (FACS). The same digestion solution was used for all organs, in order to avoid organ-specific technical confounders.
Brain. Following decapitation, the skull was cut longitudinally with scissors, and the cranium was opened with tweezers. Both brain hemispheres were carefully collected and placed into cell culture dishes containing cold PBS supplemented with 0.1% BSA (PBS + BSA). White matter was manually removed, the tissue shredded with scissors and added to a 50 ml tube containing 15 ml cold Accumax (Sigma-Aldrich) digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer (Sigma-Aldrich) followed by centrifugation at 300 g for 5 min at 4 °C. Myelin was removed by using density gradient centrifugation. Cells were recovered at the interface between a 80% Percoll layer and a 30% Percoll layer and washed in PBS + BSA to remove excess Percoll.
Caecum, large intestine, small intestine. Luminal contents were removed. The organs were cut longitudinally with scissors, rinsed several times in PBS + BSA to remove mucus, then cut into 0.5 cm pieces and placed in 50 ml tubes containing 20 ml pre-warmed (37 °C) RPMI containing 10% FCS and 5 mM EDTA (RPMI + FCS + EDTA). Samples were incubated for 25 min at 37 °C in a shaking incubator at 200 rpm. Supernatant was collected, samples were resuspended once again in 20 ml pre-warmed RPMI + FCS + EDTA, and incubated for 25 min at 37 °C in a shaking incubator at 200 rpm. These wash steps were performed to dissociate epithelial cells. After the second incubation, supernatants were collected and combined, followed by digestion of the samples in 15 ml cold Accumax for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer. Organ homogenates were combined with the epithelial cell fractions, filtered twice through a 100 µm cell strainer and washed in cold PBS + BSA.idney, lung, spleen, thymus. Organs were rinsed with cold PBS + BSA and shredded with scissors. Tissue fragments were placed into 50 ml tubes containing 20 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered through a 100 µm cell strainer. After centrifugation, 2 ml cold ACK lysis buffer (Thermo Fisher Scientific) was added for 3 min to lyse red blood cells and the reaction stopped by adding 20 ml of cold PBS + BSA. Supernatants were filtered twice through 100 µm cell strainer and washed once to remove residual ACK lysis buffer.
Liver. Three lobes were removed, rinsed with cold PBS + BSA, and shredded with scissors. Tissue fragments were placed into a 50 ml tube containing 20 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered through a 100 µm cell strainer. Hepatocytes were removed using density gradient centrifugation. Cells recovered at the interface between an 80% Percoll layer and a 30% Percoll layer were washed in PBS + BSA to remove excess Percoll.
Lymph nodes. Cervical, axillary, and inguinal lymph nodes were combined, carefully pinched with tweezers, and rinsed several times with cold PBS + BSA to release hematopoietic cells. Tissue fragments were placed into a 50 ml tube containing 10 ml cold Accumax digestion solution and incubated for 45 min at 37 °C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered twice through a 100 µm cell strainer.
Skin. Ears were harvested at the base, and the subcutaneous fat layer scrapped off with a scalpel. Tissue fragments were then shredded with scissors, placed into 50 ml tubes containing 15 ml of Accumax digestion solution, and incubated for 45 min at 37 ° C while shaking at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer and filtered twice through a 100 µm cell strainer.

Organ-specific sample collection and organ dissociation

Large intestine. The organ was removed and processed as described. Briefly, luminal contents were removed, the large intestine cut longitudinally with scissors, rinsed several times in PBS + BSA to remove mucus, then cut into 0.5 cm pieces and placed in 50 ml tubes containing 20 ml pre-warmed (37 °C) RPMI containing 10% FCS and 5 mM EDTA (RPMI + FCS + EDTA). Samples were incubated for 40 min at 37 °C in a shaking incubator at 200 rpm after. Supernatant was collected, samples resuspended once again in 20 ml pre-warmed RPMI + FCS + EDTA, and incubated for 20 min at 37 °C in a shaking incubator at 200 rpm. These wash steps were performed to dissociate epithelial cells. After the second incubation, supernatants were collected and combined, followed by incubating samples in RPMI containing 10% FCS and 15 mM HEPES (RPMI + FCS + HEPES) for 10 min at room temperature. Supernatant was discarded and samples digested in RPMI + FCS + HEPES containing 100 U/ml collagenase from Clostridium histolyticum (Sigma) for 1 hour at 37 °C in a shaking incubator at 200 rpm. Remaining tissue fragments were processed with a Dounce homogenizer. Organ homogenates were combined with the epithelial cell fractions, filtered twice through 100 µm cell strainers and washed in cold PBS + BSA.

Lung. The organ was removed and processed as described. Briefly, the organ was cut into 0.5 cm pieces and placed in gentleMACS C tubes (Miltenyi) containing 160 U/mL collagenase type 1 (Gibco) and 12 U/mL DNAse 1 (Sigma) in RPMI containing 5% FCS, and dissociated using a gentleMACS Dissociator (Miltenyi; program m_lung_01). After incubation at 37°C for 35 min in a shaking incubator at 170 rpm, digested samples were homogenized using the gentleMACS Dissociator (program m_lung_02). Subsequently, cell suspensions were filtered through 70 µM cell strainers and centrifuged for 5 min at 4 °C at 300g. After centrifugation, 1 ml cold ACK lysis buffer (Thermo Fisher) was added for 5 min to lyse red blood cells and the reaction stopped by adding 20 ml of cold PBS + BSA. Supernatants were filtered twice through 100 µm cell strainer and washed once to remove residual ACK lysis buffer.

Liver. Mice were anesthetized (ketamine/xylazine 1:3, 0.1ml/10g mouse, Vetoquinol) before cannulation of the liver and dissociation using a two-step perfusion protocol. Briefly, the liver was perfused first with 20 mL HBSS (Gibco) containing 0.5 mM EGTA (Sigma) and afterward with 20 mL L15 medium (Gibco) containing 40 mg/L liberase (Roche) at a rate of 5 mL/min. Next, the liver was removed, placed in a Petri dish with 10 ml of the same liberase-containing medium, followed by removal of the liver capsule. Hepatocytes were removed from the resulting cell suspension using density gradient centrifugation. Cells recovered at the interface between an 80% Percoll layer and a 30% Percoll layer were washed in PBS + BSA to remove excess Percoll.

Flow cytometry and FACS

Single-cell suspensions were washed once with PBS containing 0.1% BSA and 5 mM EDTA (PBS + BSA + EDTA). Cells were then incubated with anti-CD16/CD32 (clone 93, BioLegend) to prevent nonspecific binding. Single-cell suspensions were then stained with different combinations of antibodies against CD45 (PerCP-Cy5.5, clone 30-F11), TER-119 (PerCP-Cy5.5, clone TER-119), podoplanin (PE, clone 8.1.1), EpCAM (Pe-Cy7, clone 8.8), lymphotoxin beta receptor (APC, clone 5G11), CD31 (FITC, clone MEC13.3), CD90.2 (AF700, clone 30-H12), CD106 (AF647, clone 429), CD144 (BV421, clone BV13), CD324 (APC-Cy7, clone DECMA-1) (all from BioLegend), CD140a (BV605, clone APA5, BD Bioscience), MAdCAM1 (BV421, clone MECA-367, BD Bioscience), and LYVE1 (eFlour 660, clone ALY7, Thermo Fisher Scientific) for 30 min at 4 °C, followed by two washes with PBS + BSA + EDTA. Dead cells were stained either by adding Zombie Red Fixable Viability Dye or Zombie Aqua Fixable Viability Dye (both from BioLegend) immediately prior to flow cytometry characterization or cell sorting. For flow cytometry, cells were acquired with an LSRFortessa (BD Biosciences) cell analyzer using the outlined gating strategies (Extended Data Fig. 1a, 2a, 8a). For FACS, cells were sort-purified with a MoFlo Astrios (Beckman Coulter) or SH800 (Sony) using the outlined gating strategies (Extended Data Fig. 1a, 8a). Data analysis was performed using the FlowJo software (Version 10.5.3, Tree Star).

Instrument	For flow cytometry, cells were acquired with an LSRFortessa (BD Biosciences) cell analyzer. For FACS, cells were sort-purified with a MoFlo Astrios (Beckman Coulter) or SH800 (Sony).
Software	Data analysis was performed with FlowJo (Version 10.5.3 Tree Star) software.
Cell population abundance	As expected, structural cells varied in abundance across organs (Figure 1b, Extended Data Fig. 1, and Extended Data Fig. 8). It was not feasible to routinely perform post-purity checks on the sorted cell populations due to their low and varied abundance.
Gating strategy	The gating strategy is depicted in Extended Data Fig. 1a, Extended Data Fig. 2a, and Extended Data Fig. 8a.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Systematic quantitative analysis of ribosome inventory during nutrient stress

https://doi.org/10.1038/s41586-020-2446-y

Received: 2 February 2020

Accepted: 7 May 2020

Heeseon An[1,3], Alban Ordureau[1,3], Maria Körner[1,2], Joao A. Paulo[1] & J. Wade Harper[1]✉

Mammalian cells reorganize their proteomes in response to nutrient stress through translational suppression and degradative mechanisms using the proteasome and autophagy systems[1,2]. Ribosomes are central targets of this response, as they are responsible for translation and subject to lysosomal turnover during nutrient stress[3–5]. The abundance of ribosomal (r)-proteins (around 6% of the proteome; 10^7 copies per cell)[6,7] and their high arginine and lysine content has led to the hypothesis that they are selectively used as a source of basic amino acids during nutrient stress through autophagy[4,7]. However, the relative contributions of translational and degradative mechanisms to the control of r-protein abundance during acute stress responses is poorly understood, as is the extent to which r-proteins are used to generate amino acids when specific building blocks are limited[7]. Here, we integrate quantitative global translatome and degradome proteomics[8] with genetically encoded Ribo–Keima[5] and Ribo–Halo reporters to interrogate r-protein homeostasis with and without active autophagy. In conditions of acute nutrient stress, cells strongly suppress the translation of r-proteins, but, notably, r-protein degradation occurs largely through non-autophagic pathways. Simultaneously, the decrease in r-protein abundance is compensated for by a reduced dilution of pre-existing ribosomes and a reduction in cell volume, thereby maintaining the density of ribosomes within single cells. Withdrawal of basic or hydrophobic amino acids induces translational repression without differential induction of ribophagy, indicating that ribophagy is not used to selectively produce basic amino acids during acute nutrient stress. We present a quantitative framework that describes the contributions of biosynthetic and degradative mechanisms to r-protein abundance and proteome remodelling in conditions of nutrient stress.

r-Protein stoichiometry is controlled by translation and assembly mechanisms as well as by proteasomal degradation of excess r-proteins[9–11], whereas autophagy may facilitate the turnover of ribosomes en masse[5,7]. Previous studies examining the effect of amino acid withdrawal or mTOR inhibition on r-protein homeostasis in mammalian cells have primarily focused on autophagic r-protein turnover, using either immunoblotting to measure r-protein abundance for specific subunits[4] or Ribo–Keima to measure autophagic flux[5]. However, a global view of how cells regulate net ribosome abundance upon nutrient stress is lacking, as autophagy is only one of several components of the ribosome homeostasis system. To uncover the mechanisms that control r-protein levels during nutrient stress, we developed a quantitative framework for analysing r-protein abundance, synthesis, turnover and subcellular partitioning, using methods applicable to ensemble or single-cell measurements (Extended Data Fig. 1a).

We first used quantitative proteomics to examine the net balance of r-proteins after the acute withdrawal of amino acids or inhibition of mTOR with a small-molecule inhibitor, Torin1 (Tor1) (Fig. 1a–e, Extended Data Fig. 1b–d). HEK293 and HCT116 cells with or without the ATG8-conjugation (ATG5) or signalling (RB1CC1, also known as FIP200) arms of the autophagy system were subjected to amino acid withdrawal or treatment with Tor1 for 10 or 24 h, followed by 11-plex tandem mass tagging (TMT)-based proteomics (Fig. 1b). We also mined our published dataset using HEK293T (293T) wild-type, *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ cells that were subjected to the same experimental pipeline[12] (Fig. 1c, Extended Data Fig. 1b). As expected, both treatments resulted in a reduction in the levels of autophagy cargo receptors (GABARAPL2, LC3B, SQSTM1 and TEX264) and endoplasmic reticulum (ER) proteins, which was dependent on ATG5 or ATG7 and RB1CC1[12] (Extended Data Fig. 1b–e, Supplementary Table 1). By contrast, plots of the $\log_2$-transformed ratio (cells treated with nutrient stressors/untreated cells) versus the $-\log_{10}$-transformed *P* value for at least 70 out of 80 r-proteins revealed only a modest (4.6–11.6%) reduction of r-protein levels with both forms of nutrient stress; this was largely independent of autophagy (Fig. 1c–f, Extended Data Fig. 1b–d), although there was some variation in the extent of decrease between individual r-proteins (Fig. 1g, Extended Data Fig. 1f). The reduction in abundance for several r-proteins was not observable using quantitative immunoblotting (Extended Data Fig. 1g–k).

---

[1]Department of Cell Biology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [2]Present address: Department of Biochemistry, University of Würzburg, Würzburg, Germany. [3]These authors contributed equally: Heeseon An, Alban Ordureau. ✉e-mail: wade_harper@hms.harvard.edu

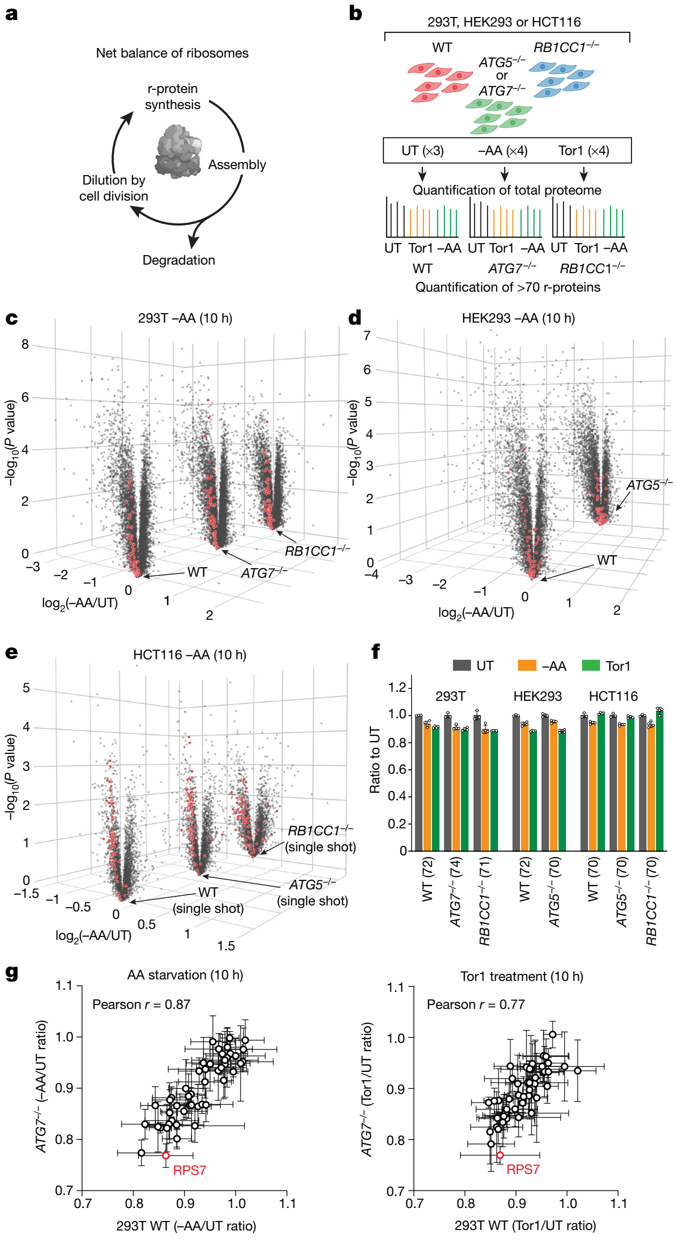


Fig. 1 | Reduction in r-protein abundance during nutrient stress is largely autophagy-independent. a, Factors affecting the net balance of ribosomes during nutrient stress. **b**, The 11-plex TMT pipeline to measure ribosome abundance during nutrient stress with or without active autophagy. Normalized total cell extracts were processed for eight 11-plex TMT mass spectrometry (TMT-MS) experiments. –AA, withdrawal of amino acids; UT, untreated; WT, wild type. **c**–**e**, Volcano plots of the $-\log_{10}$-transformed *P* value versus the $\log_2$-transformed ratio of –AA cells/UT cells for 293T cells (WT, $n = 8{,}029$ proteins; $ATG7^{-/-}$, $n = 8{,}373$ proteins; $RB1CC1^{-/-}$, $n = 8{,}332$ proteins) (**c**), HEK293 cells (WT, $n = 7{,}531$ proteins; $ATG5^{-/-}$, $n = 7{,}504$ proteins) (**d**) or HCT116 cells (WT, $n = 3{,}779$ proteins; $ATG5^{-/-}$, $n = 3{,}761$ proteins; $RB1CC1^{-/-}$, $n = 3{,}671$ proteins) (**e**). In **c**–**e**, $n = 3$ (untreated) or $n = 4$ (–AA) biologically independent samples. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 1. **f**, Mean ratio value (±s.d.) measured for at least 70 r-proteins treated as in **b** (numbers of proteins are shown in parentheses; $n = 3$ (untreated) or $n = 4$ (–AA and Tor1) biologically independent samples). The 293T data are from a previous study[12]. **g**, Relative abundance of individual r-proteins in 293T cells after either 10 h of amino acid withdrawal (left) or 10 h of treatment with Tor1 (right). For every condition, 48 r-proteins with an error range of less than ±10% are plotted. Data are mean ± s.d. for $n = 3$ (untreated) or $n = 4$ (–AA and Tor1) biologically independent samples. See also Extended Data Fig. 1, Supplementary Table 1.

Ribo–Halo analysis of r-proteins

The absence of an obvious autophagy-dependent loss of r-proteins (on the basis of proteomics with normalization for the total protein input) created a paradox, as ribophagic flux has been observed with similar treatments using Ribo–Keima reporters[5]. Changes in r-protein abundance could be a result of multiple mechanisms, including translational inhibition, degradation or effects on ribosome dilution owing to reduced cell division (Fig. 2a). To simultaneously examine the synthesis of new r-proteins, the fate of pre-existing r-proteins and the dilution of ribosomes through cell division, we fused a Halo cassette with the surface-exposed C termini of endogenous *RPS3* and *RPL29* genes in HCT116, 293T or HEK293 cells (hereafter called Ribo–Halo) (Extended Data Fig. 2a, b). Halo-fused proteins can be covalently labelled in a temporally controlled manner using distinct fluorescent ligands, which facilitates the analysis of single cells by flow cytometry or imaging and ensembles of cells using in-gel fluorescence. Halo tagging did not affect the abundance or translation rates of r-proteins, or the response to mTOR signalling (Extended Data Fig. 2c, d). We also optimized concentrations, wash conditions and compensation protocols to ensure complete Halo labelling with individual red (tetramethylrhodamine; TMR) or green (R110) fluorescent ligands (Extended Data Fig. 2e–i; see Methods).

We first measured the effect of mTOR inhibition (14 h) on the total abundance of RPS3–Halo and RPL29–Halo in single HCT116 cells by flow cytometry using a 1-h pulse-chase-labelling experiment with a cell-permeable red Halo ligand (Fig. 2b). The mean fluorescence per cell was reduced by around 25% in both cell lines after Tor1 treatment (Fig. 2c, d). In parallel, we found that treatment with Tor1 led to a reduction of around 10% in cell diameter (about a 25% reduction in volume)[13], as well as a reduction in the rates of cell division, thereby maintaining the density of RPS3 and RPL29 molecules in single cells (Extended Data Fig. 2j, k). Thus, changes in the number and size of cells probably underlie the discrepancy between the effects of nutrient stress measured by Ribo–Halo and those measured by total proteomics, as the latter method uses a normalized total protein input.

To simultaneously measure the dilution of pre-existing r-proteins and the synthesis of new r-proteins, HCT116 Ribo–Halo cells were treated for 1 h with red Halo ligand to irreversibly label pre-existing r-proteins, free ligand was rapidly washed out and cells were then incubated in rich medium with green Halo ligand for 8, 16 or 24 h to label newly translated r-proteins (Fig. 2e, Extended Data Fig. 2l). As expected, the relative contribution of pre-existing RPS3 and RPL29 to the entire pool of these proteins in single cells decreased with time, reflecting dilution by cell division (Fig. 2e, Extended Data Fig. 2l). Fifty percent of pre-existing r-proteins were observed at around 16 h, in accordance with independently measured rates of cell proliferation (Extended Data Fig. 2k). The total abundance of r-proteins per cell was maintained by the synthesis of new r-proteins (labelled with green Halo ligand), consistent with a doubling in ribosome number during one cell cycle. After inhibition of mTOR during labelling with green Halo ligand (Fig. 2f, Extended Data Fig. 2m), we observed a reduced dilution of pre-existing r-proteins at 8, 16 and 24 h in parallel with reduced cell division, and a reduction in the production of newly synthesized (green) r-proteins—apparently reflecting translational inhibition. A reduction in the total number of r-proteins (red and green) by 10–30% was consistent with single-ligand labelling. In addition, a reduction in the synthesis of new r-proteins and the dilution of pre-existing r-proteins was observed using imaging or in-gel fluorescence in Ribo–Halo cells after treatment with Tor1 (14 h) (Fig. 2g, h, Extended Data Fig. 2n–q). Consistent with residual mTOR activity in HCT116 cells relative to 293T cells, we observed differential repression of new RPS3–Halo by pulse-chase Halo labelling (40% versus 60%) (Extended Data Fig. 2o, r). Notably, the abundance, dilution and new synthesis of r-proteins, and

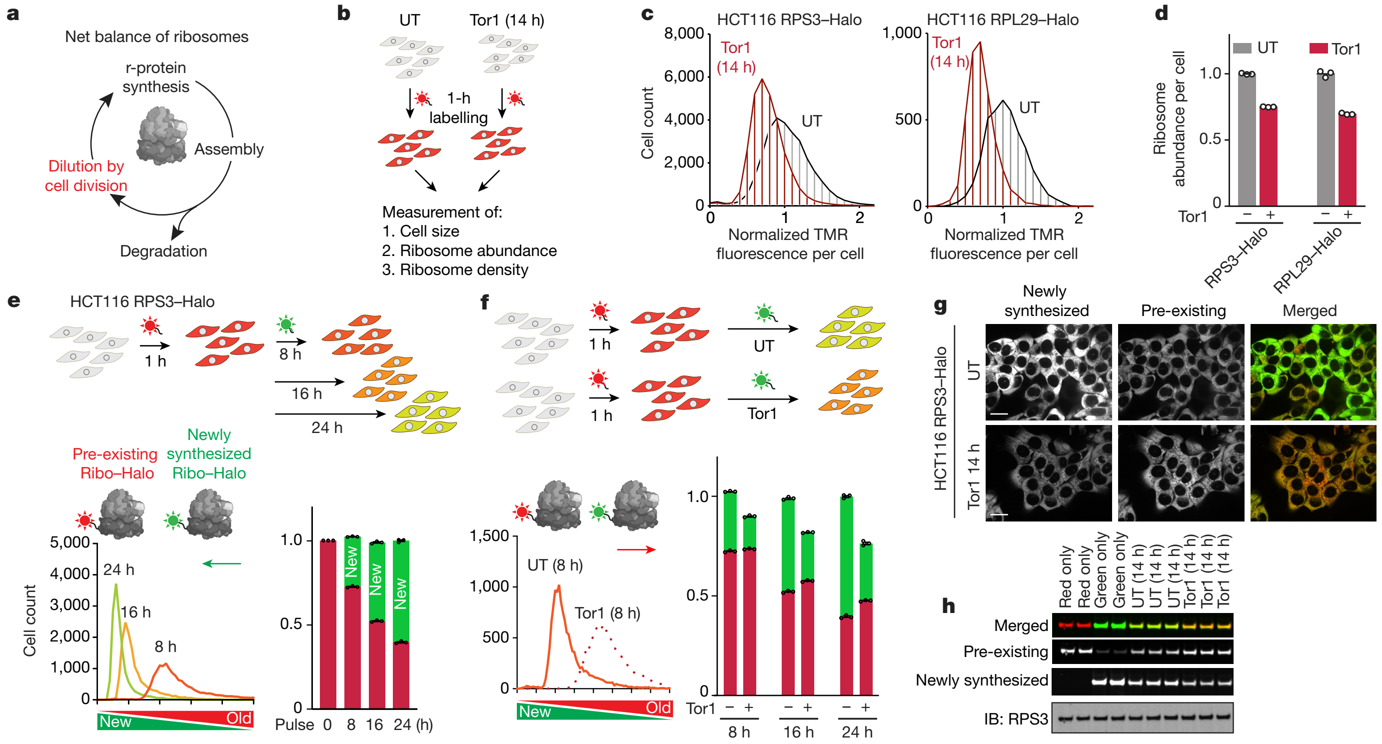


Fig. 2 | Density, synthesis and dilution of r-proteins in single cells using Ribo–Halo. a, Contribution of dilution by cell division to the net ribosome balance during nutrient stress. **b**, Measurement of r-protein concentration in single cells using one-colour Halo labelling. **c**, Normalized TMR signal in RPS3–Halo cells (>3×10^5) and RPL29–Halo cells (>4×10^3) incubated with or without 200 nM Tor1 (14 h), followed by 1 h TMR ligand treatment and flow cytometry. Representative of three independent experiments. **d**, Mean ± s.d. of triplicate data from **c**. **e**, Pulse-chase Ribo–Halo in HCT116 cells. Pre-existing RPS3–Halo was labelled with red TMR ligand (1 h) and newly synthesized RPS3–Halo was labelled with green R110 ligand for 8, 16 or 24 h (top) before flow cytometry (frequency histogram, bottom left; around 1.4×10^4 cells analysed). Bottom right, mean values (± s.d.) from the triplicate experiments. See Methods for details. **f**, Top, scheme for two-colour Ribo–Halo r-protein biogenesis and dilution labelling with or without Tor1. Pre-existing RPS3–Halo was labelled with TMR ligand (1 h), and newly synthesized RPS3–Halo with or without Tor1 (200 nM) was labelled with green R110 ligand. Bottom left, ratio of the 620-nm emission (red) intensity to the 550-nm emission (green) intensity plotted against cell populations. Bottom right, mean values (± s.d.) from the triplicate experiments for the 8 h, 16 h and 24 h pulse chases. **g**, Imaging of pre-existing red TMR-labelled RPS3–Halo and newly synthesized green R110-ligand-labelled RPS3–Halo with or without Tor1 (200 nM, 14 h). Scale bars, 20 μm. **h**, In-gel fluorescence of RPS3–Halo as in **f**. Gels were then transferred for immunoblotting (IB) with α-RPS3. Experiments in **g**–**h** were repeated more than three times with similar results. The full immunoblot is shown in Extended Data Fig. 11f. See also Extended Data Figs. 2, 3. For gel source data, see Supplementary Fig. 1.

the changes in cell size after inhibition of mTOR (14 or 24 h), were unaltered in cells lacking ATG7 or RB1CC1 (Extended Data Fig. 3a–n). Thus, the overall abundance of r-proteins in response to mTOR inhibition reflects a reduction in the synthesis of new r-proteins and a reduced dilution of pre-existing r-proteins, the magnitudes of which may vary with cell type. Autophagic turnover of r-proteins and ribosomal RNA (rRNA) is unlikely to provide a mechanism to reverse translational suppression and the reduction in cell growth under conditions of acute stress.

Translatomics during nutrient stress

We next developed a quantitative translatome approach to systematically examine r-protein synthesis during nutrient stress using metabolic labelling with azidohomoalanine (AHA) (Fig. 3a, b). Activation of AHA by $\mathrm{tRNA}^{\mathrm{Met}}$-synthetase allows the incorporation of azides into proteins[14] that can then be functionalized through click chemistry[8], thereby allowing the measurement of ongoing translation using TMR alkyne for in-gel quantification or biotin alkyne for purification with streptavidin resin and TMT-based proteomics (AHA-TMT) (Fig. 3b). We optimized AHA labelling (250 μM) to minimize effects on mTOR activity while allowing sufficient incorporation of AHA (Extended Data Fig. 4a–g, Methods). As expected on the basis of ribosome profiling and the incorporation of ^{35}S-labelled Cys or Met[1], overall translation was reduced by 60–70% after inhibition of mTOR or withdrawal of amino acids (3 h), as measured by TMR labelling (Fig. 3c, Extended Data Fig. 4h). We then quantified the translation of more than 8,200 proteins by AHA-TMT proteomics. The vast majority of proteins were translationally repressed after inhibition of mTOR or withdrawal of amino acids (3 h)—a notable exception being eIF4EBP1 after mTOR inhibition[2,15] (Fig. 3d, e, Extended Data Fig. 4i). The majority of r-proteins were translationally repressed beyond the median value, consistent with 5′ terminal oligopyrimidine motifs in their mRNAs[1]. Inhibition of mTOR repressed the translation of r-proteins more selectively than did the withdrawal of amino acids, despite similar proteome-wide levels of translational suppression (Extended Data Fig. 4h, j). In addition, we identified proteins for which translation was more selectively suppressed (group 1), largely unchanged (group 2) or differentially affected by the two treatments (groups 3 and 4) (Extended Data Fig. 4k–o). Notably, the extent of translational suppression in *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ 293T cells was comparable to that in wild-type cells (Fig. 3e, Extended Data Fig. 4p–r). Thus, global translational suppression of r-proteins contributes to the control of ribosome abundance during acute nutrient stress; this is not discernibly affected by the ability of cells to recycle amino acids through autophagy in acute conditions; and, finally, the extent of translational suppression varies between individual r-proteins (Extended Data Fig. 4s).

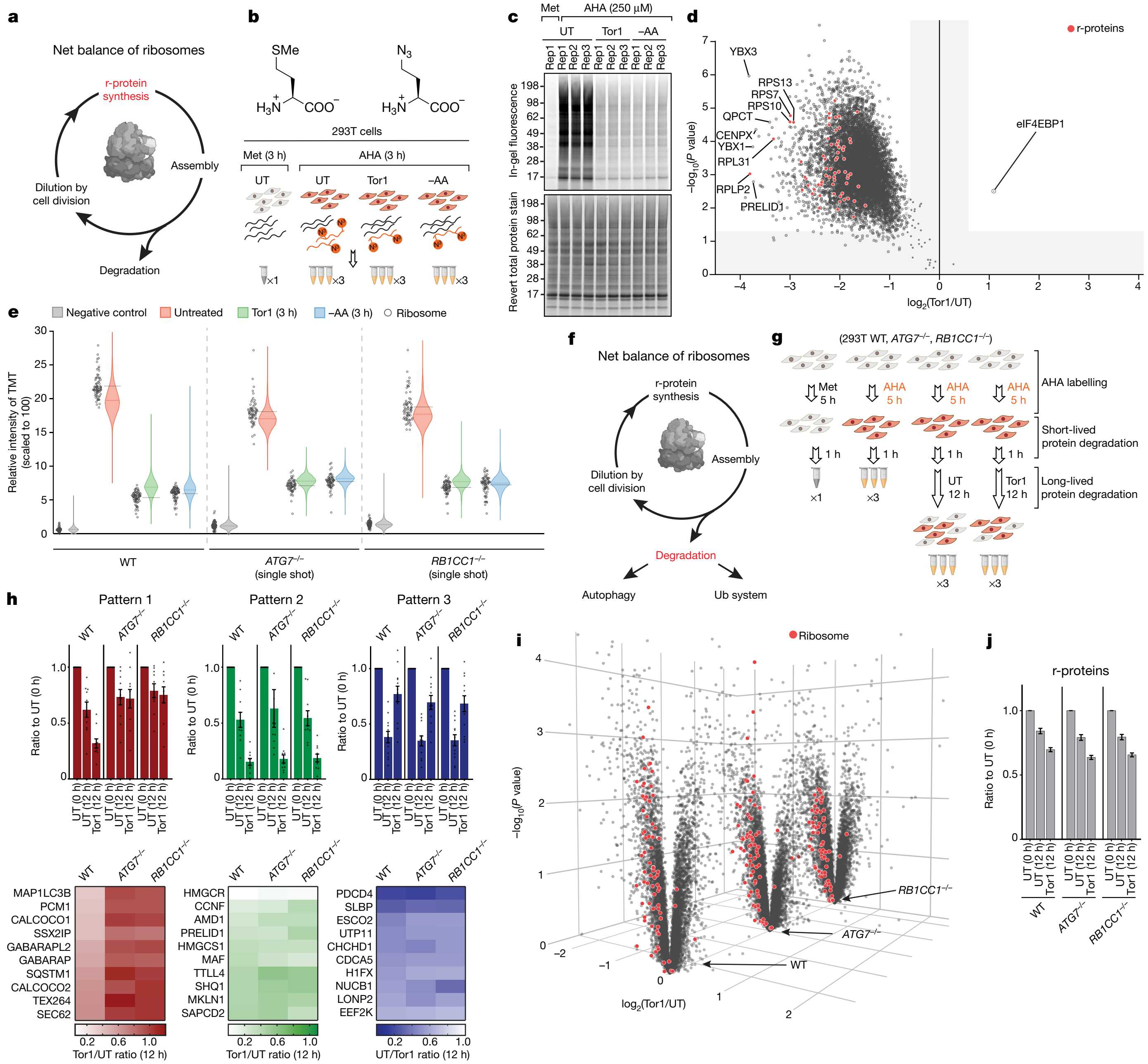


Fig. 3 | Global translatome and degradome analysis during nutrient stress. **a**, Contribution of r-protein synthesis to the net ribosome balance during nutrient stress. **b**, AHA-based translatomics to measure translational suppression during nutrient stress. The final steps are as follows: (1) western blot to check sample quality; (2) In-gel TMR fluorescence analysis; (3) biotin–streptavidin enrichment; and (4) TMT 10-plex mass spectrometry (MS) analysis. **c**, Extracts from 293T cells with or without amino acid withdrawal or Tor1 treatment in the presence of Met or AHA (250 μM, 3 h each) were labelled with TMR and analysed by in-gel fluorescence ($n = 3$ biologically independent samples for UT, −AA and Tor1; $n = 1$ for Met). **d**, Extracts from 293T cells (as in **b**) were clicked with biotin before streptavidin enrichment and proteomics. Plot shows the $-\log_{10}$-transformed *P* value versus the $\log_2$-transformed ratio of Tor1-treated cells/UT cells (3 h each) for quantification ($n = 8{,}285$ proteins; $n = 3$ biologically independent samples). *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 2. **e**, Violin plots for TMT intensity of newly synthesized proteins in wild-type, *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ 293T cells ($n = 8{,}285$, $n = 3{,}336$ and $n = 3{,}418$ proteins, respectively). Black horizontal lines show the median abundance of r-proteins (black circles); coloured horizontal dotted lines show the median of all proteins. Violin plots represent the distribution and density of the whole dataset (centre line, median; limits, minimum and maximum values). **f**, Contribution of r-protein degradation to the net ribosome balance during nutrient stress. Ub, ubiquitination. **g**, Schematic of degradomics analysis using AHA pulse labelling (5 h) to examine proteome turnover and autophagy dependence during mTOR inhibition (12 h). The final steps are as follows: (1) western blot to check sample quality; (2) click with biotin alkyne; (3) streptavidin enrichment; (4) TMT 10-plex MS analysis. **h**, Pattern 1, accelerated degradation via autophagy; pattern 2, accelerated degradation independent of autophagy; pattern 3, stabilization independent of autophagy. Mean ratio values (± s.e.m.) of 10 representative proteins ($n = 3$ biological triplicate experiments). **i**, Volcano plots of the $-\log_{10}$-transformed *P* value versus the $\log_2$-transformed ratio of Tor1-treated cells/UT cells (12 h each) for 293T wild type, *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ cells ($n = 8{,}304$, $n = 8{,}319$ and $n = 8{,}590$ proteins, respectively; $n = 3$ biologically independent samples). *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 3. **j**, Mean ratio values (± s.e.m.) for AHA-labelled r-proteins as in **g** ($n = 58$ r-proteins, s.d. < 0.3 filter for every condition; $n = 3$ biologically independent samples). See also Extended Data Figs. 4–6, Supplementary Tables 2, 3. For gel source data, see Supplementary Fig. 1.

Global degradomics during nutrient stress

To directly examine the degradation of r-proteins during nutrient stress and the contribution of autophagy to turnover (Fig. 3f), we pulse-labelled wild-type, *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ 293T cells with AHA, allowed 1 h for the turnover of short-lived proteins in rich medium and then used TMT-based proteomics to compare the abundance of biotinylated AHA-labelled proteins with or without mTOR inhibition (12 h) (Fig. 3g, Extended Data Fig. 5a, b). Among the proteins quantified (more than 8,300), we found three patterns of turnover (Fig. 3h, Extended Data Fig. 5c, d): pattern 1, autophagy-dependent degradation, including several autophagy receptors[16]; pattern 2, autophagy-independent turnover that is enhanced by mTOR inhibition, including several proteins that are known to be degraded by the proteasome upon mTOR inhibition[17]; and pattern 3, autophagy-independent stabilization upon mTOR inhibition. Notably, r-proteins conformed to pattern 2, and were degraded in a largely autophagy-independent manner—although degradation rates varied for individual r-proteins with a correlated rank order with or without Tor1 (R^2 value of around 0.7) (Fig. 3i, j, Extended Data Fig. 5e, f). As expected[12], 326 ER proteins were degraded in an autophagy-dependent manner, and serve as a positive control for this degradomics approach to detect autophagy substrates (Extended Data Fig. 5g–i). To confirm these results, we performed a time-course experiment to examine global degradomics at 5, 10 and 15 h after inhibition of mTOR (Extended Data Fig. 6a–g). Consistently, mTOR inhibition induced a faster turnover of r-proteins in a time-dependent manner, which was largely independent of an active autophagy pathway.

Proteome partitioning during amino acid stress

Given that several ribosome assembly steps occur in the nucleus or nucleolus, we examined whether nutrient stress alters the nuclear-to-cytosolic ratio of r-proteins (Extended Data Fig. 7a). After optimizing isolation of nuclei and cytosol (Extended Data Fig. 7b–f), we quantified cytosolic and nuclear protein and phosphoprotein pools with or without the withdrawal of amino acids (3 h) (Extended Data Fig. 7g–l). We observed a decrease in the nuclear abundance of FOXK1, a decrease in the phosphorylation of RPS6, FOXK1, FOXK2 and CAD, and an increase in the nuclear abundance of TFEB, MITF and TFE3—indicative of mTOR inhibition[18] (Extended Data Fig. 7j–m). We also identified several nucleolar ribosome-biogenesis factors (PWP1, SDAD1 and NVL)[19] that accumulate in the cytosol after the withdrawal of amino acids (Extended Data Fig. 7j, k, n–p). By contrast, the distribution of r-proteins was largely unchanged after amino acid withdrawal. As expected, around 91% of 60S r-proteins and 95% of 40S r-proteins are cytoplasmic in untreated cells—consistent with the longer time required to assemble the 60S subunit at steady state and the larger fraction of 40S subunits that are assembled in the cytosol[20,21] (Extended Data Fig. 8a). After the withdrawal of amino acids, the nuclear pool of r-proteins was reduced by 20% (around 1.5% of the total ribosome pool) (Extended Data Fig. 8b–d), indicating that the redistribution of r-proteins has a very minor role in ribosome homeostasis in these conditions.

Ribophagic flux during nutrient stress

Our finding that cells that lack the capacity for autophagy have no obvious defects in the response to nutrient stress, on the basis of total-proteome, Ribo–Halo, translatome and degradome analysis, led us to further quantify ribophagic flux. Using Ribo–Keima reporters[5], we estimated that 3–4% of ribosomes are degraded in the lysosome after Tor1 treatment (10 h) (Extended Data Fig. 8e–h), consistent with previous studies[5]. The magnitude of autophagic flux is such that RPS3 turnover as measured by AHA-TMT proteomics (around 20% (s.d. < 0.1) in 12 h) is not obviously sensitive to the deletion of *ATG7* or *RB1CC1* (Extended Data Fig. 8i). NUFIP1—a subunit of the nucleoplasmic R2TP pathway that is involved in 60S rRNA biogenesis[2]—was reported to traffic to the cytosol in response to nutrient stress and to function as a selective ribophagy receptor for lysosomal delivery[4,22]. The absence of NUFIP1 re-localization after amino acid withdrawal (as measured by TMT-based proteomics; Extended Data Fig. 7j, k) led us to re-evaluate its role in ribophagy. First, *NUFIP1*$^{-/-}$ 293T cells exhibit reduced levels of r-proteins in rich medium, in line with the previously defined role of R2TP in 60S biogenesis (Extended Data Fig. 9a, b). Second, *NUFIP1* deletion had no effect on the total Ribo–Halo abundance, synthesis or dilution (Extended Data Fig. 9c–e), or on r-protein abundance as measured by TMT-based proteomics (Extended Data Fig. 9f–i) during mTOR inhibition. Third, RPS3–Keima flux was unaffected in *NUFIP1*$^{-/-}$ cells subjected to nutrient stress (Extended Data Fig. 9j). We conclude that NUFIP1 is not required for ribophagic flux in these stress conditions.

Single amino acid or purine depletion

The enrichment (around twofold) of basic amino acids in r-proteins relative to the proteome, together with the fact that ribosomes are rich in rRNA, has led to the hypothesis that ribosomes are selectively used for autophagy in response to decreased levels of these building blocks[4,23,7]. We used our r-protein analysis pipeline to assess the effects of selective withdrawal of Arg, Lys or Leu on ribosome homeostasis and ribophagic flux (Fig. 4a). Compared with Tor1 treatment or complete starvation of amino acids, selective withdrawal of Leu or Arg resulted in only a partial loss of mTOR activity (Extended Data Fig. 10a, b), as described previously[24]. However, withdrawal of Arg resulted in stronger translational inhibition, as measured by puromycin incorporation assay (Extended Data Fig. 10b, c), two-colour RPS3–Halo assay (Fig. 4b, Extended Data Fig. 10d) and translatome analysis (Fig. 4c, d, Extended Data Fig. 10e–h). The order of translational suppression was Arg ≥ Tor1 > Lys > Leu for both the total proteome and r-proteins. However, ribophagic flux[5] was less than 3.4% in all cases (Extended Data Fig. 10i), indicating that acute loss of basic amino acids does not selectively promote ribophagy.

We also found that 6-mercaptopurine (6-MP), an inhibitor of hypoxanthine-guanine phosphoribosyltransferase 1 (HPRT1) that blocks production of inosine monophosphate (IMP) and guanosine monophosphate (GMP)[25] (Extended Data Fig. 11a), markedly suppressed proteome and r-protein translation (Extended Data Fig. 11b–d), and also suppressed both dilution and new synthesis in the two-colour Ribo–Halo assay (Extended Data Fig. 11e, f). However, the reduction in r-protein abundance measured by TMT-based proteomics was ATG5-independent (Extended Data Fig. 11g). Consistent with this, ribophagic flux as measured using the RPS3–Keima processing assay was not detectably induced during treatment with 6-MP (24 h), in stark contrast to treatment with arsenite[5] (Extended Data Fig. 11h), indicating that autophagy is unlikely to be a major driver in the regulation of r-protein abundance in response to purine depletion.

Framework for r-protein homeostasis

Here we have developed an experimental pipeline for broadly probing the integration of biosynthetic and degradative mechanisms that control signal-dependent proteome remodelling, with a focus on r-proteins. We found that the major contributing factors in the control of r-protein abundance in conditions of either mTOR inhibition or amino acid withdrawal were: (1) translational suppression of r-proteins; (2) reduced dilution of pre-existing ribosomes, reflecting lower rates of cell proliferation; and (3) non-autophagic degradation of r-proteins (Fig. 4e). By contrast, the contribution of ribophagy to overall changes in ribosome abundance was relatively small (around 3–4% in 10–12 h). The finding that r-protein turnover parallels that of proteasomal targets by degradomics (pattern 2, Fig. 3h) suggests that the proteasome is involved, but relevant ubiquitin ligases[7] remain unknown.

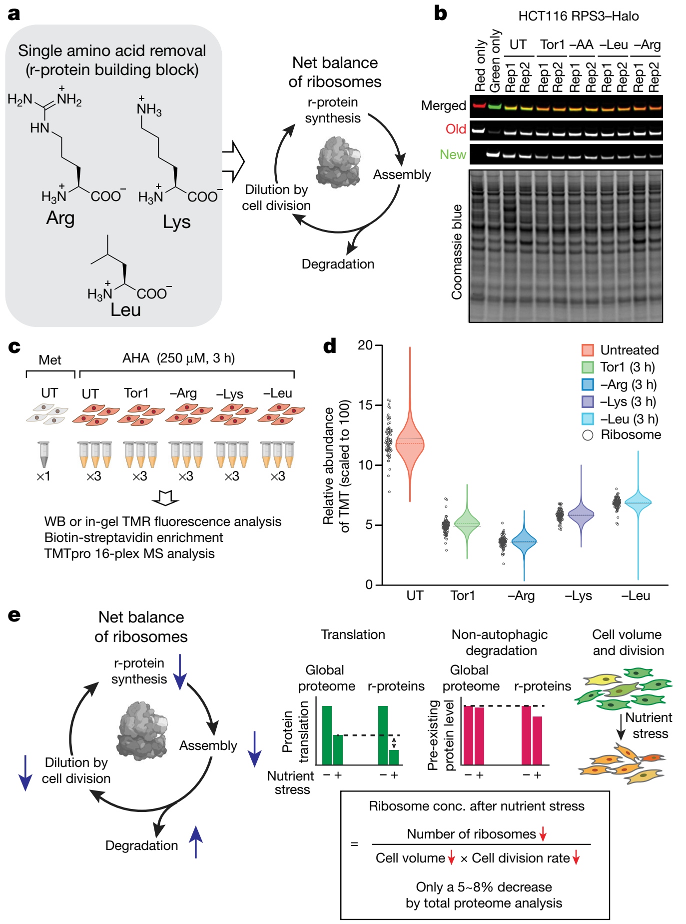


Fig. 4 | Analysis of r-protein homeostasis in response to single amino acid perturbations. a, Points of intersection of individual amino acids with net ribosome production. **b**, HCT116 RPS3–Halo cells were left untreated, subjected to the withdrawal of all amino acids or the selective withdrawal of Leu or Arg, or incubated with Tor1 before analysis of pre-existing and newly synthesized RPS3–Halo as in Fig. 2f. For quantification of $n = 2$ biologically independent experiments, see Extended Data Fig. 10d. **c**, Experimental workflow for AHA translatome analysis of 293T cells with or without the indicated amino acids or with Tor1 treatment. **d**, Violin plots showing the relative proteome translation rates and the effect of removal of the indicated amino acids or inhibition of mTOR on the translation of individual r-proteins (black circles). Black horizontal lines show the median abundance of r-proteins; coloured horizontal dotted lines show the median of all proteins. Violin plots represent the distribution and density of the whole dataset (centre line, median; limits, minimum and maximum values). $n = 3{,}372$ quantified proteins (single shot). **e**, Ribosome homeostasis framework. Left, points of regulation. Blue arrows indicate the net effect of nutrient stress. Middle two panels, comparison of the effects of nutrient stress on translation and degradation (primarily through non-autophagic mechanisms) of the global proteome and r-proteins. Far right, effect of nutrient stress on the dilution of r-proteins through cell division. The concentration of ribosomes (ribosome conc.) in a population of cells after nutrient stress will reflect a decrease in the number of r-proteins through translational and degradative mechanisms, corrected for by changes in cell volume and cell division. See also Extended Data Fig. 10, Supplementary Table 6. For gel source data, see Supplementary Fig. 1.

We developed a quantitative framework (Fig. 4e, Extended Data Fig. 11i, j) that facilitates an understanding of the factors that contribute to the cellular concentration of ribosomes, taking into account the number of ribosomes, timing of cell division and cell volume. A central element of the model is that the number of ribosomes per cell increases by twofold over the course of the cell cycle and is reset upon division. In conditions of nutrient stress, changes in the rates of r-protein synthesis and degradation slow the increase in r-protein abundance (numerator), which is partially compensated for by changes in cell volume and the timing of division (denominator) (Extended Data Fig. 11i, j). Incorporation of our data on translational inhibition, degradation, cell volume and cell division rate into this model accurately estimates the total change in ribosome concentration measured by total proteomics. However, our present model uses zero-order kinetics for changes in translation, turnover and cell division, and more-precise models will require measurements of rates of change during the stress interval. Our framework highlights the importance of considering multiple mechanisms of r-protein homeostasis simultaneously to fully understand similarities and differences across cell systems that may respond to nutrient stress in distinct ways. In this regard, ribophagy in budding yeast has been proposed to be a major contributor to proteome remodelling in response to nitrogen deficiency[3]. This apparent differential use of autophagy in yeast compared to cultured mammalian tissue cells could reflect the fact that r-protein abundance in yeast is an order of magnitude higher than that seen in mammalian cells; however, it is also complicated by the fact that the duration of nutrient stress applied (24 h) is around 16 times the cell division time in yeast[7,26]. Notably, approximately 10% of ER proteins are degraded within 10 h of the withdrawal of amino acids in human cells, and this is blocked by genetic ablation of autophagy[12], as verified here by degradomics (Extended Data Fig. 5g–i). Given that the contribution of ER proteins to the total proteome mass rivals that of r-proteins[6,7], the ER may constitute a major target for autophagic degradation in response to the withdrawal of amino acids or inhibition of mTOR. Ribosomes may also be degraded to some extent by autophagy through association with the rough ER, as visualized in situ by electron microscopy[27]. It seems probable that in vivo and with chronic nutrient stress, autophagy will be responsible for the generation of amino acids such as arginine, as has been shown in the context of tumorigenesis[28].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2446-y.

1. Thoreen, C. C. et al. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485**, 109–113 (2012).
2. Saxton, R. A. & Sabatini, D. M. mTOR signaling in growth, metabolism, and disease. *Cell* **168**, 960–976 (2017).
3. Kraft, C., Deplazes, A., Sohrmann, M. & Peter, M. Mature ribosomes are selectively degraded upon starvation by an autophagy pathway requiring the Ubp3p/Bre5p ubiquitin protease. *Nat. Cell Biol.* **10**, 602–610 (2008).
4. Wyant, G. A. et al. NUFIP1 is a ribosome receptor for starvation-induced ribophagy. *Science* **360**, 751–758 (2018).
5. An, H. & Harper, J. W. Systematic analysis of ribophagy in human cells reveals bystander flux during selective autophagy. *Nat. Cell Biol.* **20**, 135–143 (2018).
6. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13**, 3497–3506 (2014).
7. An, H. & Harper, J. W. Ribosome abundance control via the ubiquitin-proteasome system and autophagy. *J. Mol. Biol.* **432**, 170–184 (2020).
8. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proc. Natl Acad. Sci. USA* **103**, 9482–9487 (2006).
9. Sung, M. K. et al. A conserved quality-control pathway that mediates degradation of unassembled ribosomal proteins. *eLife* **5**, e19105 (2016).
10. Nguyen, A. T. et al. UBE2O remodels the proteome during terminal erythroid differentiation. *Science* **357**, eaan0218 (2017).
11. Yanagitani, K., Juszkiewicz, S. & Hegde, R. S. UBE2O is a quality control factor for orphans of multiprotein complexes. *Science* **357**, 472–475 (2017).
12. An, H. et al. TEX264 is an endoplasmic reticulum-resident ATG8-interacting protein critical for ER remodeling during nutrient stress. *Mol. Cell* **74**, 891–908 (2019).
13. Miettinen, T. P. & Björklund, M. Mevalonate pathway regulates cell size homeostasis and proteostasis through autophagy. *Cell Rep.* **13**, 2610–2620 (2015).

14. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl Acad. Sci. USA* **99**, 19–24 (2002).
15. Thoreen, C. C. The molecular basis of mTORC1-regulated translation. *Biochem. Soc. Trans*. **45**, 213–221 (2017).
16. Kirkin, V. & Rogov, V. V. A diversity of selective autophagy receptors determines the specificity of the autophagy pathway. *Mol. Cell* **76**, 268–285 (2019).
17. Zhao, J., Zhai, B., Gygi, S. P. & Goldberg, A. L. mTOR inhibition activates overall protein degradation by the ubiquitin proteasome system as well as by autophagy. *Proc. Natl Acad. Sci. USA* **112**, 15790–15797 (2015).
18. Settembre, C. et al. TFEB links autophagy to lysosomal biogenesis. *Science* **332**, 1429–1433 (2011).
19. Liu, Y. et al. PWP1 mediates nutrient-dependent growth control through nucleolar regulation of ribosomal gene expression. *Dev. Cell* **43**, 240–252 (2017).
20. de la Cruz, J., Karbstein, K. & Woolford, J. L. Jr. Functions of ribosomal proteins in assembly of eukaryotic ribosomes in vivo. *Annu. Rev. Biochem*. **84**, 93–129 (2015).
21. Kakihara, Y. & Houry, W. A. The R2TP complex: discovery and functions. *Biochim. Biophys. Acta* **1823**, 101–107 (2012).
22. Shim, M. S., Nettesheim, A., Hirt, J. & Liton, P. B. The autophagic protein LC3 translocates to the nucleus and localizes in the nucleolus associated to NUFIP1 in response to cyclic mechanical stress. *Autophagy* https://doi.org/10.1080/15548627.2019.1662584 (2019).
23. Wolfson, R. L. & Sabatini, D. M. The dawn of the age of amino acid sensors for the mTORC1 pathway. *Cell Metab*. **26**, 301–309 (2017).
24. Darnell, A. M., Subramaniam, A. R. & O'Shea, E. K. Translational control through differential ribosome pausing during amino acid limitation in mammalian cells. *Mol. Cell* **71**, 229–243 (2018).
25. Hoxhaj, G. et al. The mTORC1 signaling network senses changes in cellular purine nucleotide levels. *Cell Rep*. **21**, 1331–1346 (2017).
26. Weinberg, D. E. et al. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep*. **14**, 1787–1799 (2016).
27. Chino, H., Hatta, T., Natsume, T. & Mizushima, N. Intrinsically disordered protein TEX264 mediates ER-phagy. *Mol. Cell* **74**, 909–921 (2019).
28. Poillet-Perez, L. et al. Autophagy maintains tumour growth through circulating arginine. *Nature* **563**, 569–573 (2018).

Methods

Cell lines

HEK293 (human embryonic kidney, fetus, ATCC CRL-1573, RRID: CVCL_0045), HCT116 (human colorectal carcinoma, male, ATCC CCL-247, RRID: CVCL_0291) and HEK293T (human embryonic kidney, fetus, ATCC CCL-3216, RRID: CVCL_0063) cells were grown in Dulbecco's modified Eagle's medium (DMEM, high glucose and pyruvate) supplemented with 10% fetal calf serum and maintained in a 5% CO_2 incubator at 37 °C. Karyotyping (GTG-banded karyotype) of HCT116, HEK293 and 293T cells was performed by Brigham and Women's Hospital Cytogenomics Core Laboratory. All cell lines were found to be free of mycoplasma using the Mycoplasma Plus PCR assay kit (Agilent).

Generation of RPS3–HaloTag7 and RPL29–HaloTag7 knock-in cell lines using CRISPR–Cas9 gene editing

Guide RNAs (gRNAs) targeting the C-terminal region of human *RPS3* and *RPL29* genes were designed using the CHOPCHOP website (http://chopchop.cbu.uib.no/). The guide sequences for *RPS3* gene (5′- GACATACCTGTTATGCTGTG-3′) or *RPL29* gene (5′-GAGATATCTCTGCCAACATG-3′) were assembled into a pX459 plasmid[29]. Donor vectors were constructed by assembling a HaloTag7 transgene with upstream and downstream homology arms (650 nucleotide each) into a digested pSMART plasmid by Gibson assembly. HEK293, HEK293T and HCT116 cells were transfected with donor and gRNA vectors (1 to 1 ratio) by Lipofectamine 3000 (Invitrogen). Five days after the transfection, the pool of transfected cells was treated with 100 nM Halo-TMR ligand for 1h, followed by washing three times. Fluorescence-positive cells were sorted into 96-well plates by flow cytometry (MoFlo Astrios EQ, Beckman Coulter). Three weeks later, the expanded single-cell colonies were screened for the integration of the HaloTag7 transgene by immunoblotting with α-RPS3 or α-RPL29, followed by genotyping.

Generation of gene-knockout cell lines using CRISPR–Cas9 gene editing

RB1CC1, *ATG5*, *ATG7* and *NUFIP1* knockout in 293T, HEK293 and HCT116 cell lines was carried out by plasmid-based transfection of Cas9/gRNA using a pX459 plasmid as described[29]. First, six gRNAs for *ATG5*, *ATG7* and *NUFIP1* and four guide RNAs for *RB1CC1* knockout were designed using the CHOPCHOP website. Puromycin selection was performed 24 h after the transfection, and 48 h after the transfection, the gene-cleavage efficiency of each guide RNA was measured by Surveyor assay. The following gRNAs were shown to have the best cutting efficiency among the tested guides: 5′- GTCCAAGGCACTACTAAAAG-3′ (exon 2) for *ATG7*; 5′-GATCACAAGCAACTCTGGAT-3′ (exon 5) for *ATG5*; 5′-GAAGAATCTGGGCGTCGAA-3′ (exon 1) for *NUFIP1*; and 5′-GCTACGATTGACACTAAAGA-3′ (exon 7) for *RB1CC1*. Single cells were sorted into 96-well plate using a limiting dilution method, and expanded clonal cells were screened by immunoblotting with α-ATG7, α-ATG5, α-RB1CC1, NUFIP1 and α-LC3B antibodies.

Reagents

Antibodies. The following antibodies were used: RPS3 (Cell Signaling Technology, 9538); RPL28 (Abcam, ab138125), RPS15a (Bethyl Lab, A304-990A-T), ATG7 (Cell Signaling Technology, 8558S), Keima (MBL international, M182-3), LC3B (MBL international, M186-3), ATG5 (Cell Signaling Technology, 12994), p70 S6K phospho-T389 (Cell Signaling Technology, 9234S), phospho-S6 ribosomal protein Ser253, 236 (Cell Signaling Technology, 4858), TEX264 (Sigma, HPA017739), RPL23 (Bethyl Lab, A305-010A-T), RPL7 (Bethyl Lab, A300-741A-T), RPL29 (Proteintech Group, 15799-1-AP), Tubulin (Abcam, ab7291), SQSTM1 (Novus Biologicals, H00008878-M01), anti-puromycin antibody (EMD millipore, MABE343), NUFIP1 (Proteintech Group, 12515-1-AP), RPS6 (Cell Signaling Technology, 2217), 4EBP1 (Cell Signal Technology, 9644), Lamin A/C (Cell Signal Technology, 4777), TFEB (Cell Signal Technology, 4042), SDAD1 (Bethyl Lab, A304-692A-T), NVL (Proteintech Group, 16970-1-AP), c-Maf (RnD systems, MAB8227-SP), IRDye 800CW Streptavidin (LI-COR, 926-32230), IRDye 800CW goat anti-rabbit IgG H+L (LI-COR, 925-32211) and IRDye 680 RD goat anti-mouse IgG H+L (LI-COR, 926-68070).

Chemicals, peptides and recombinant proteins. The following chemicals, peptides and recombinant proteins were used: Tor1 (Cell Signal Technology, 14379), SAR405 (APExBio, A8883), azidohomoalanine (Click Chemistry Tools, 1066-1000), 5-TAMRA alkyne (Click Chemistry Tools, 1255-1), biotin-PEG4-alkyne (Click Chemistry Tools, TA105-25), HaloTag R110Direct Ligand (Promega, G3221), HaloTag TMR (5mM) (Promega, G8251), 6-MP monohydrate (Sigma, 852678-1G-A), sodium ascorbate (VWR international, 95035-692), poly-L-lysine solution (Sigma, P4832), FluoroBrite DMEM (Thermo Fisher Scientific A, 1896701), benzonase nuclease HC (Millipore, 71205-3), urea (Sigma, Cat#U5378), sodium dodecyl sulfate (SDS) (Bio-Rad, Cat#1610302), Surveyor Mutation Detection Kit (Integrated DNA Technologies, 706025), Revert Total Protein Stain kit (LI-COR, P/N926-11010), 1,1,1,3,3,3-hexafluoro-2-propanol (Sigma, 52517), DMEM, high-glucose, pyruvate (Gibco/Invitrogen, 11995), DMEM, low-glucose, without amino acids (US Biological, D9800-13), TCEP (Gold Biotechnology), puromycin (Gold Biotechnology, P-600-100), formic acid (Sigma-Aldrich, 94318), protease inhibitor cocktail (Sigma-Aldrich, P8340), PhosSTOP (Sigma-Aldrich, 4906845001), trypsin (Promega, V511C), Lys-C (Wako Chemicals, 129-02541), Rapigest SF Surfactant (Glixx Laboratories, Cat#GLXC-07089), EPPS (Sigma-Aldrich, Cat#E9502), 2-chloroacetamide (Sigma-Aldrich, C0267), TMT 11plex Label Reagent (Thermo Fisher Scientific, Cat#90406 & #A34807), TMTpro 16plex Label Reagent (Thermo Fisher Scientific, Cat#A44520), hydroxylamine solution (Sigma Cat#438227), Empore SPE Disks C18 (3M - Sigma-Aldrich Cat#66883-U), Sep-Pak C18 Cartridge (Waters Cat#WAT054960 and #WAT054925), SOLA HRP SPE Cartridge, 10 mg (Thermo Fisher Scientific, Cat#60109-001), High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific, Cat#84868), High-Select Fe-NTA Phosphopeptide Enrichment Kit (Thermo Fisher Scientific, Cat#A32992), Bio-Rad Protein Assay Dye Reagent Concentrate (Bio-Rad, #5000006) and Pierce Quantitative Colorimetric Peptide Assay (Thermo Fisher Scientific, #23275).

Preparation of amino-acid-free medium

DMEM powder (4.16 g, US biological) and 1.85 g sodium bicarbonate were dissolved in 400 ml H_2O. To the solution, 1.75 g of glucose, 5 ml of 100× sodium pyruvate (final concentration 1 mM) and 50 ml of dialysed fetal calf serum were added, and 150 µl of 6N HCl was slowly added to bring the pH to 7.2. The final volume was adjusted to 500 ml. The medium was filtered through a 0.2-µm filter and kept at 4 °C.

Preparation of medium without Met, Lys or Arg

One litre of base medium for the indicated amino acid withdrawal was prepared by dissolving DMEM powder (8.32g, US biological), 3.7 g/l sodium bicarbonate, and 3.5 g/l glucose, 30 mg/l glycine, 63 mg/l cystine 2HCl, 580 mg/l glutamine, 42 mg/l histidine HCl-H_2O, 105 mg/l isoleucine, 66 mg/l phenylalanine, 42 mg/l serine, 95 mg/l threonine, 16mg/ltryptophan,104mg/ltyrosine2Na2H_2O,94mg/lvalineand105mg/l leucine in 880 mL H_2O, and the pH was titrated to pH 7.2 using 2M HCl. The base medium was filtered through a 0.2-µm filter and kept at 4 °C. Before each experiment, 0.9 equivalence (v/v) of the base medium was added by 0.1 equivalence (v/v) of dialysed bovine serum albumin, 0.01 equivalence (v/v) of 100× sodium pyruvate (final concentration 1 mM) and 0.001 equivalence of 1000× amino acids (methionine stock: 37.5 mg/ml (final 37.5 mg/l = 250 µM), lysine HCl stock: 146 mg/ml (final 146 mg/l), arginine HCl stock: 84 mg/ml (final 84 mg/l) except the limiting amino acid.

Preparation of leucine-free medium with or without methionine preparation

The solubility of leucine in H_2O was too low to be added at the last step as a concentrate. Therefore, the leucine-free medium had to be prepared separately from the −Lys and −Arg media. To prepare 1 l of medium, DMEM powder (8.32 g, US biological), 3.7g/l sodium bicarbonate and 3.5 g/l glucose, 30 mg/l glycine, 63 mg/l cystine 2HCl, 580 mg/l glutamine, 42 mg/l histidine HCl-H_2O, 105 mg/l isoleucine, 66 mg/l phenylalanine, 42 mg/l serine, 95 mg/l threonine, 16 mg/l tryptophan, 104 mg/l tyrosine 2Na $2H_2O$, 94 mg/l valine, 146 mg/l Lys, and 84 mg/l Arg were dissolved in 880 ml H_2O, and the pH was titrated to pH 7.2 using 2M HCl. The base medium was filtered through a 0.2-µm filter and kept at 4 °C. Before each experiment, 0.9 equivalence (v/v) of the base medium was added by 0.1 equivalence (v/v) of dialysed bovine serum albumin, 0.01 equivalence of 100× sodium pyruvate (v/v) and 0.001 equivalence (v/v) of 1000× methionine (final 37.5 mg/l = 250 µM) or 0.01 equivalence (v/v) of 100× AHA (final 250 µM).

Cell lysis and immunoblotting assay

Immunoblotting assay was performed based on a previously reported method[5]. In brief, cells were cultured in the presence of the corresponding stress to around 40–50% confluency in 6-well plates, 10-cm or 15-cm dishes. After removing the medium, the cells were washed with DPBS three times, then in-house RIPA buffer (50 mM HEPES, 150 mM NaCl, 1% sodium deoxycholate, 1% NP-40, 0.1% SDS, 2.5 mM $MgCl_2$, 10 mM sodium glycerophosphate, 10 mM sodium biphosphate) containing mammalian protease inhibitor cocktail (Sigma), Phos-STOP, and 20 unit/ml benzonase (Millipore) were added directly onto the cells. Cell lysates were collected by cell scrapers and sonicated on ice three times, followed by centrifugation (13,000 rpm, 5 min). The concentration of the supernatant was measured by Bradford assay, and the whole cell lysate was further denatured by the addition of LDS sample buffer supplemented with 50 mM DTT, followed by boiling at 75 °C for 5 min. Twenty or thirty micrograms of each lysate was loaded onto the 4–20% Tris-Glycine gel (Thermo Fisher Scientific) or 4–12% NuPAGE Bis-Tris gel (Thermo Fisher Scientific), followed by SDS–PAGE with Tris-Glycine SDS running buffer (Thermo Fisher Scientific) or MES SDS running buffer (Thermo Fisher Scientific), respectively. The proteins were electro-transferred to PVDF membranes (0.45 µm, Millipore), and then the total protein was stained by Revert total protein stain kit (LI-COR) or Ponceau staining (Thermo Fisher Scientific). The membrane was then blocked with 5% non-fat milk (room temperature, 30 min), incubated with the indicated primary antibodies (4 °C, overnight), washed three times with TBST (total 30 min), and further incubated either with fluorescent IRDye 680RD goat anti-mouse IgG H+L, IRDye 800RD goat anti-mouse IgG H+L, or IRDye 800CW goat anti-rabbit IgG H+L secondary antibody (1:15,000) for 1 h. After thorough wash with TBST for 30 min, near infrared signal was detected using OdysseyCLx imager and quantified using ImageStudioLite (LI-COR). For quantitative immunoblotting of endogenously tagged Ribo–Keima reporter cells, at least 70 µg of total lysate was loaded onto an SDS–PAGE gel owing to the low level of processed Keima. A 4–20% Tris-Gly gel was used to resolve the proteins.

Flow cytometry analysis for one-colour Ribo–Halo labelling

The corresponding Ribo–Halo cells were plated onto 24-well plates one day before the nutrient stress. The cells were left untreated or treated with 200 nM Tor1 for the indicated time. One hour before collecting the cells, 100 nM of TMR ligand was added in addition to the Tor1 for HaloTag labelling. The cells were washed with fresh DMEM with or without Tor1, three times with 10-min duration each. After trypsin treatment, the collected cells were resuspended in 250 µl FACS buffer (1× DPBS, 1 mM EDTA, 1% FBS, 25 mM HEPES, final pH 7.3–7.5) and analysed by flow cytometry (LSR-II Analyser, BD Biosciences). The data were processed by FlowJo software.

Flow cytometry analysis for two-colour Ribo–Halo labelling

The corresponding Ribo–Halo cells were plated onto 24-well plates, and 100 nM of TMR-Halo ligand was added and incubated for 1 h to label the pre-existing ribosomes, then washed away. The cells were incubated with 2 ml of fresh DMEM for 10 min in the dark to remove the remaining free TMR ligand. This was repeated for three times in total. Then the cells were grown either in rich medium or in Tor1 treatment medium, both containing 50 nM R110 Halo ligand. In parallel, control Ribo–Halo cells were stained only with TMR ligand or with R110 ligand, the signal of which represented 100% ribosome abundance and was used as a normalization factor. The cells were than collected after trypsin treatment, resuspended in 250 µl FACS buffer (1× DPBS, 1 mM EDTA, 1% FBS, 25 mM HEPES, final pH 7.3–7.5), and analysed by flow cytometry (LSR-II Analyser, BD Biosciences). The data were processed by FlowJo software. In brief, the 488- and 561-nm intensities of individual cells (>3,000 cells) were exported to Microsoft Excel. Each fluorescence signal from the single cells was normalized by the signal intensity from the 100% red- or 100% green-ligand-treated cells for the same duration. The old to new ribosome ratio distribution graph was processed by dividing the normalized red signal with the normalized green signal coming from the same single cells and plotted using Prism software.

In-gel fluorescence analysis for two-colour Ribo–Halo labelling

Related to Figs. 2, 4. The corresponding Ribo–Halo cells were plated onto 12-well plates, and 100 nM of TMR-Halo ligand[30] was added and incubated for 1h to label the pre-existing ribosomes, then washed away. The cells were incubated with 2 ml of fresh DMEM for 10 min in the dark to remove the remaining free TMR ligand. This was repeated for three times in total. Then the cells were grown either in rich medium or in Tor1 treatment medium, both containing 50 nM R110 Halo ligand. In parallel, control Ribo–Halo cells were stained only with TMR ligand or with R110 ligand, the signal of which represented 100% ribosome abundance and was used as a normalization factor. The cells were than collected after trypsin treatment, resuspended in RIPA buffer and sonicated three times on ice. Following Bradford assay, 15 µg of the lysates were taken from each sample and resolved by SDS–PAGE using a 4–12% NuPAGE Bis-Tris gel (Thermo Fisher Scientific). The in-gel TMR and R110 signals were detected by ChemiDoc MP (Bio-Rad). The gel was then either stained with Coomassie blue staining for loading control, or the proteins were transferred to a PVDF membrane for the subsequent immunoblotting analysis. The ratio change of the TMR and R110 signals was analysed using Image Lab software (Bio-Rad).

Measurement of cell size

Related to Extended Data Figs. 2, 3. Cells were grown on a 24-well plate with or without Tor1 (200 nM) for 14 h. Then, the cells were treated with trypsin for 3 min, followed by the addition of 1 ml DMEM. The cells were collected and spun down at 1,000 rpm for 3 min, and the DMEM was removed. The cell pellet was resuspended in 500 µl of DMEM with or without Tor1 (200 nM) at room temperature. A 75-µl quantity of each sample (around 2.0×10^5 cells/ml) was injected to Moxi Cassette and the size measured by Moxi GO II. The single cell-size measurements were processed by FlowJo and plotted using Prism software.

Confocal microscopy

Related to Fig. 2. Live-cell confocal microscopy was performed based on a previously reported method[5]. In brief, cells were plated onto a 35-mm glass-bottom dish (no. 1.5, 14-mm glass diameter, MatTek) pre-treated with poly-L-lysine. Two days later, the cells were incubated in 100 nM TMR-ligand-containing medium for 1 h. Following the washing steps (three times for 10 min each), the cells were incubated with or without Tor1 in the presence of 50 nM R110 Halo ligand for 14 or 24 h. The medium was changed to FluoroBrite DMEM (Thermo Fisher Scientific) before the live-cell imaging. The cells were imaged using a Yokogawa

CSU-X1 spinning disk confocal with Spectral Applied Research Aurora Borealis modification on a Nikon Ti motorized microscope equipped with a Nikon Plan Apo 60×/1.40 N.A or 100×/1.40 N.A objective lens. Pairs of images for ratiometric analysis of TMR and R110 fluorescence were collected sequentially using 100 mW 488-nm and 100 mW 561-nm solid state lasers attenuated and controlled with an AOTF (Spectral Applied Research LMM-5), and the emission was collected with a 525/50-nm or 620/60-nm filter (Chroma Technologies), respectively. Wide-field fluorescence images of Hoeschst were collected using a Lumencor SOLA light source, 395/35-nm excitation and 480/40-nm emission filters (Chroma Technologies). Both confocal and wide-field images were acquired with the same Hamamatsu ORCA-ER cooled CCD camera and MetaMorph software. For the analysis, the same gamma, brightness and contrast were applied for each image using Fiji software.

AHA incorporation and TMR-in-gel fluorescence analysis for translatome analysis

Related to Fig. 3. Wild-type 293T cells were plated onto a 10-cm dish with around 20% confluency 24 h before the experiment. Commercial DMEM medium was replaced with in-house prepared DMEM medium containing either 250 μM Met or 250 μM AHA in addition to the indicated nutrient stress medium. The cells were washed three times with DPBS and lysed by the addition of RIPA buffer and sonication (4 °C, three times). Following Bradford assay, 30 μg of lysate was subjected to immunoblotting for pS6K, pS6 and 4E-BP1 for a quality control of the proper inhibition of mTOR. From the remaining lysates, an equal protein amount was collected across the samples, reduced by 5 mM TCEP (10 min, room temperature), then alkylated by chloroacetamide (20 mM final concentration, 15 min, room temperature). $CHCl_3$/MeOH precipitation was performed to remove any remaining AHA in the cell lysates. The white protein disk was then resuspended in 2% SDS (50 mM HEPES 150 mM NaCl, pH 7.2, 2.5 mM TCEP) and sonicated once with a tip sonicator. Ten microlitres of the resuspended lysate was added with 10 μl of 2× Click-master mix containing 2 mM $CuSO_4$, 2 mM sodium ascorbate, 200 μM TBTA ligand and 200 μM TMR-alkyne in H_2O. The click reaction mixture was then incubated in the dark for 1 h (with rocking, room temperature; final sample concentration 0.825 μg/μl). The mixture was then resolved by SDS–PAGE, and the TMR signal detected by ChemiDoc MP (Bio-Rad) and analysed by Image Lab software (Bio-Rad).

AHA incorporation, biotin-click and streptavidin enrichment for translatome analysis

Related to Fig. 3. Wild-type 293T cells were plated onto a 10-cm dish with around 20% confluency 24 h before the experiment. Commercial DMEM medium was replaced with in-house prepared DMEM medium containing either 250 μM Met or 250 μM AHA in addition to the indicated nutrient stress. The cells were washed three times with DPBS and lysed by the addition of RIPA buffer and sonication (three times). Following Bradford assay, 30 μg of the lysate was subjected to immunoblotting for pS6K, pS6,and 4E-BP1 for a quality control of the proper inhibition of mTOR. We also confirmed that AHA addition (250 μM) without prior Met withdrawal minimized the effects on mTOR activity using pS6K band intensity[31]. From the remaining lysates, an equal protein amount was collected across the samples, then reduced by 5 mM TCEP (10 min, room temperature) and alkylated by chloroacetamide (20 mM final concentration, 15 min, room temperature). $CHCl_3$/MeOH precipitation was performed to remove any remaining AHA in the cell lysates. The white protein disk was then resuspended in 2% SDS (50 mM HEPES 150 mM NaCl, pH 7.2, 2.5 mM TCEP) and sonicated once with a tip sonicator. At least 1 mg of the lysates resuspended in 2% SDS was taken and reacted with the click reagents (final concentration: 1 mM $CuSO_4$, 1 mM sodium ascorbate, 100 μM TBTA ligand, and 100 μM biotin-alkyne; 1% SDS), and the mixture was incubated for 2 h with rocking at room temperature. This was followed by $CHCl_3$/MeOH precipitation again to remove excess biotin-alkyne, and the lysate was resuspended with 200 μl of 2% SDS (TCEP 2.5 mM) and diluted with RIPA to make <1 mg/ml concentration of the lysate (final SDS concentration is <0.5%). Meanwhile, 10 μl of high-capacity streptavidin agarose beads multiplied by the number of samples (for example, 100 μl for 10 samples; the bead has a 10 mg/ml BSA-biotin capture capacity. Therefore, 10 μl of beads can roughly capture 100 μg of biotinylated proteins when the average molecular weight is assumed to be 66 kDa) were washed with 1 ml RIPA and distributed to 10 × 1.5 ml Eppendorf tubes. The lysates were added to the beads and incubated overnight at room temperature. Flow-through was then stored for the quality control, and the beads were washed with RIPA ×2, 1 M KCl, 0.1 M Na_2CO_3, 2 M urea in HEPES buffer ×2 and RIPA ×2. The beads were transferred onto hydrophilic PTFE membrane filter cups and washed with water twice. Completely dried beads were then resuspended in 50 μl of hexafluoroisopropanol (HFIP), incubated for 5 min with shaking and eluted by spinning. Fifty microlitres of HFIP was added to the beads again, and the eluates were combined and dried in a speedvac for 10 min (the boiling point of HFIP is 58 °C; volatile). This sample was used for either western blotting or TMT-MS analysis. Note: as a quality control, we gathered the flow-through and blotted it for streptavidin-IRDye800 for equal capture. Also, the flow-through was precipitated with the $CHCl_3$/MeOH method followed by click reaction with TMR-alkyne (an orthogonal alkyne reagent to the biotin-alkyne) to check the efficiency of the click reaction. On the basis of this method, we concluded that the efficiency of the initial click reaction was always over 90% when compared with the negative control that does not contain AHA.

AHA incorporation and TMT-MS sample preparation for degradome analysis

Related to Fig. 3. Equal numbers of the wild-type 293T cells were plated onto a 10-cm dish with around 20% confluency 24 h before the experiment. Commercial DMEM medium was replaced with in-house prepared DMEM medium containing either 250 μM Met or 250 μM AHA and the cells were incubated for 5 h for AHA incorporation. The AHA-containing medium was removed, and the cells were washed with rich DMEM. The cells were incubated in the rich DMEM supplemented with 10% FBS for 1 h for the degradation of short-lived proteins with AHA incorporation. Then the control cells were collected immediately by addition of the RIPA buffer (containing 0.5% SDS instead of 0.1%), whereas the untreated and nutrient-stress samples were further grown in rich medium or Tor1 (200 nM)-containing medium for 12 h. To avoid the dilution factor of AHA labelled proteome through cell division, we collected the lysates from the entire dish and processed for quality control immunoblotting without normalizing the total proteome level using Bradford assay. The remaining lysates were reduced by 5 mM TCEP (10 min, room temperature), alkylated by chloroacetamide (20 mM final concentration, 15 min, room temperature), and proteins were precipitated by the $CHCl_3$/MeOH method. The white protein disk was then resuspended in 2% SDS (50 mM HEPES 150 mM NaCl, pH7.2, 2.5 mM TCEP) and sonicated once with a tip sonicator. A small aliquot of the lysates was reacted with TMR alkyne for a quality check, and the rest was clicked with biotin alkyne followed by streptavidin capture. In this case, the beads were washed with more stringent buffers (RIPA ×2, 2% SDS ×2, 3M urea ×2, 0.1 M Na_2CO_3, RIPA ×2, ddH_2O ×3).

Preparation of nuclear–cytosolic partitioning samples for immunoblotting and TMT-MS analysis

Related to Extended Data Fig. 7. The final protocol used in this study (Method 3 in Extended Data Fig. 7b) was modified from a previous study[32]. HEK 293T cells were plated onto 10-cm dishes with around 20% confluency 24 h before the experiment. The cells were either left untreated or treated with −AA medium for three hours, then washed three times with DPBS. Then the cells were lysed by adding 800 μl of the lysis buffer (50 mM HEPES pH 7.2, 150 mM NaCl, 0.1%

NP-40, 10 mM glycerophosphate, 10 mM sodium biphosphate, protease inhibitor cocktail, 2.5 mM MgCl$_2$) directly on to the dish, and the lysates were collected by scraping and pipetted up and down 6 times. Over 90% of the cells showed cell-membrane rupture with an intact nucleus when observed under the microscope after Trypan blue staining. Fifty microlitres was taken and snap-frozen using liquid nitrogen (whole-cell lysate). The rest of the lysate was spun down at 7,000 rpm for 30 s and the supernatant taken and snap-frozen (cytosolic fraction). The remaining pellet was washed with 1 ml DPBS and spun again at 7,000 rpm for 30 s, and the supernatant was removed (the pellet is the nuclear fraction). The nuclear-fraction pellet was resuspended in 300 µl RIPA buffer containing benzonase and sonicated for three times. The cytosolic fraction was added with 80 µl of 10× RIPA buffer to equalize the surfactant concentration as in the nuclear fraction, and sonicated. Bradford assay was performed using the whole-cell lysate, cytosolic fractions and the nuclear fraction. One hundred micrograms of the proteins from both nuclear and cytosolic fractions were taken and processed for MS analysis, and 15 µg of each fraction was taken and subjected to immunoblotting as quality control. This method resulted in around 3.2–3.6 times more loading of the nuclear fractions compared to the cytosolic fractions when calculated back to the total protein level. This ratio was applied for the final calculation of the nuclear–cytosolic protein abundance in the TMT-MS analysis.

Methods in Extended Data Fig. 7b: Method 1: (a) Buffer: 1% TritonX, 10 mM β-glycerophosphate, 10 mM sodium pyrophosphate, 40 mM HEPES pH 7.4, 2.5 mM MgCl$_2$, protease inhibitor cocktail. (b) Lysis method: the cell lysates were incubated for 15 min at 4 °C with rocking, followed by centrifugation for 5 min at 13,000 rpm. The pellet was washed with DPBS once and resuspended in RIPA buffer for nuclear fraction. (c) Total time, around 25 min. Method 2: (a) Buffer: 0.1% TritonX, 10 mM β-glycerophosphate, 10 mM sodium pyrophosphate, 40 mM HEPES pH 7.4, 2.5 mM MgCl$_2$, protease inhibitor cocktail. (b) Lysis method: the cell lysates were incubated for 15 min at 4 °C with rocking, followed by centrifugation for 5 min at 13,000 rpm. The pellet was washed with DPBS once and resuspended in RIPA buffer for nuclear fraction. (c) Total time, around 25 min. Method 3: (a) Buffer: 0.1% NP40 in DPBS (1 mM KH$_2$PO$_4$, 150 mM sodium chloride, 5.6 mM Na$_2$HPO$_4$, protease inhibitor cocktail, pH 7.3–7.5 (b) Lysis method: the lysates gathered in the lysis buffer were pipetted three times and centrifuged at 13,000 rpm for 10 s. The pellet was washed with DPBS once and resuspended in RIPA buffer for nuclear fraction. (c) Total time, around 3 min. Method 4: (a) Buffer: 0.05% NP40, 10 mM HEPES, 1.5 mM MgCl$_2$, 10 mM KCl, 0.5 mM DTT, pH around 7.3. (b) Lysis method: the cells were collected in the lysis buffer and sat on ice for 10 min for osmotic cell lysis, followed by centrifugation at 3,000 rpm for 10 min, 4 °C. The supernatant was snap-frozen while the pellet was washed with the lysis buffer, and centrifuged for 1 min at 3,000 rpm. The pellet was resuspended in RIPA buffer for nuclear fraction. (c) Total time, around 25 min.

Proteomics workflow

An extensive description of proteomics methods and detailed parameters is included in the first sheet of each Supplementary Table.

Sample preparation. Samples were reduced and alkylated, chloroform-methanol precipitated, reconstituted in 100 mM EPPS (pH 8.5) and digested by Lys-C and then by trypsin. Samples were TMT-labelled for 60 min at room temperature. After a labelling efficiency check, samples were quenched, pooled and desalted for subsequent LC–MS/MS analysis. When indicated, pooled sample was first offline fractionated with basic pH reversed-phase liquid chromatography in a 96-well plate and combined for a total of 24 fractions[33] before desalting and subsequent liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis.

Data acquisition. Samples were analysed on Orbitrap Tribrid Series mass spectrometers coupled to a Proxeon EASY-nLC pump (Thermo Fisher Scientific). Peptides were separated on a 35-cm column packed in-house using a 95 to 110 min gradient. MS^1 data were collected using the Orbitrap (120,000 resolution). MS^2 scans were performed in the ion trap with CID fragmentation (isolation window 0.7 Da; rapid scan; NCE 35%). Each analysis used the Multi-Notch MS^3-based TMT method[34], to reduce ion interference compared to MS^2 quantification, combined in some instance with newly implemented Real Time Search analysis software[35,36], and with the FAIMS Pro Interface (using previously optimized 3 CV parameters for TMT multiplexed samples[37]). MS^3 scans were collected in the Orbitrap using a resolution of 50,000 and NCE of 65 (TMT) or 45 (TMTpro) for HCD fragmentation. The closeout was set at two peptides per protein per fraction, so that MS^{3}s were no longer collected for proteins having two peptide-spectrum matches (PSMs) that passed quality filters[36].

Data analysis. Mass spectra were processed using a Comet-based (v.2018.01 rev.2) in-house software pipeline[38,39] or Sequest-HT using Proteome Discoverer (v.2.3.0.420 - Thermo Fisher Scientific). Database searching included all canonical entries from the human reference proteome database (UniProt Swiss-Prot - 2019-01) and sequences of common contaminant proteins. Searches were performed using a 20 ppm precursor ion tolerance, and recommended product ion parameters for ion trap MS/MS were used. TMT tags on lysine residues and peptide N termini (+229.163 Da for amino-TMT or +304.2071 Da for TMTpro) and carbamidomethylation of cysteine residues (+57.021 Da) were set as static modifications, and oxidation of methionine residues (+15.995 Da) was set as a variable modification. For phosphopeptide analysis, +79.9663 Da was set as a variable modification on serine, threonine and tyrosine residues. PSMs were filtered to a 1% false discovery rate (FDR) using linear discriminant analysis as described previously[38]. Using the Picked FDR method[40], proteins were filtered to the target 1% FDR level. Phosphorylation site localization was determined using the AScore algorithm[41]. For reporter ion quantification, a 0.003 Da window around the theoretical *m*/*z* of each reporter ion was scanned, and the most intense *m*/*z* was used. Reporter ion intensities were adjusted to correct for the isotopic impurities. Peptides were filtered to include only those peptides with a sufficient summed signal-to-noise ratio across all TMT channels. An isolation purity of at least 0.7 (70%) in the MS^1 isolation window was used for samples analysed without online real-time searching. For each protein, the filtered peptide TMT or TMTpro signal-to-noise values were summed to create protein quantification values[38].

Protein quantification values were exported for further analysis in Microsoft Excel and Perseus[42] and statistical tests and parameters used are indicated in the corresponding Supplementary Tables. In brief, two-way Welch's *t*-test analysis was performed to compare two datasets, using s0 parameter (in essence a minimum fold-change cut-off) and correction for multiple comparisons was achieved by the permutation-based FDR method, both functions that are built-in in Perseus software. For whole-cell proteome analysis, each reporter ion channel was summed across all quantified proteins and normalized assuming equal protein loading of all samples. For the AHA-derived translatome, as well as nuclear–cytosolic fractionation, no normalization based on loading was performed. For AHA-derived degradome normalization, we used a previously reported strategy[43], which is based on the assumption that there are stable proteins within the pool of proteins measured, the amounts of which decay very little during the time course of the experiment.

Supplementary Tables list all quantified proteins as well as associated TMT reporter intensity and ratio change to control channels used for quantitative analysis.

Annotations for bona fide organellar protein markers were assembled using the proteins that had scored with confidence 'very high' or 'high' from the previously published HeLa dataset[44]. Annotations for ribosome and autophagy components were manually curated and assembled based on the literature. Nuclear annotations are based on the dataset for this paper (Supplementary Table 4).

Statistics and reproducibility

All statistical data were calculated using GraphPad Prism 7 or Perseus. Comparisons of data were performed by two-way ANOVA with Tukey's multiple comparisons test; *P* values of less than 0.01 were considered significant. All experiments were repeated at least three times unless otherwise indicated.

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All the mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE repository (http://www.proteomexchange.org/): dataset 1 (related to Supplementary Table 1; PXD017852, PXD017853); dataset 2 (related to Supplementary Table 2; PXD018252); dataset 3 (related to Supplementary Table 3; PXD017857, PXD018158); dataset 4 (related to Supplementary Table 4; PXD017856, PXD017855); dataset 5 (related to Supplementary Table 5; PXD017858, PXD017851); dataset 6 (related to Supplementary Table 6; PXD017861, PXD017860, PXD017859). Source data are provided with this paper. Full gel data for immunoblots are provided in Supplementary Fig. 1. All datasets generated within this study are available online, and the reagents are available from the corresponding author on request. Source data are provided with this paper.

29. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
30. Ohana, R. F. et al. HaloTag7: a genetically engineered tag that enhances bacterial expression of soluble proteins and improves protein purification. *Protein Expr. Purif.* **68**, 110–120 (2009).
31. Gu, X. et al. SAMTOR is an *S*-adenosylmethionine sensor for the mTORC1 pathway. *Science* **358**, 813–818 (2017).
32. Suzuki, K., Bose, P., Leong-Quong, R. Y., Fujita, D. J. & Riabowol, K. REAP: a two minute cell fractionation method. *BMC Res. Notes* **3**, 294 (2010).
33. Paulo, J. A. et al. Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 *S. cerevisiae* proteins across 10 carbon sources. *J. Proteomics* **148**, 85–93 (2016).
34. McAlister, G. C. et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).
35. Erickson, B. K. et al. Active instrument engagement combined with a real-time database search for improved performance of sample multiplexing workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).
36. Schweppe, D. K. et al. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
37. Schweppe, D. K. et al. Characterization and optimization of multiplexed quantitative analyses using high-field asymmetric-waveform ion mobility mass spectrometry. *Anal. Chem.* **91**, 4010–4016 (2019).
38. Huttlin, E. L. et al. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
39. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
40. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* **14**, 2394–2404 (2015).
41. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol*. **24**, 1285–1292 (2006).
42. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
43. McShane, E. et al. Kinetic analysis of protein stability reveals age-dependent degradation. *Cell* **167**, 803–815 (2016).
44. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016).

Acknowledgements This work was supported by the National Institutes of Health (grants R37NS083524, RO1AG011085 and RO1GM095567 to J.W.H. and grant RO1GM132129 to J.A.P.). We acknowledge the Nikon Imaging Center (Harvard Medical School) for imaging assistance. We thank S. Gruver and M. Kirschner for providing MoxiGo II access and training.

Author contributions H.A., A.O. and J.W.H. conceived the study. H.A., A.O. and M.K. performed all experiments. J.A.P. provided mass spectrometry expertise. The paper was written by H.A., A.O. and J.W.H.

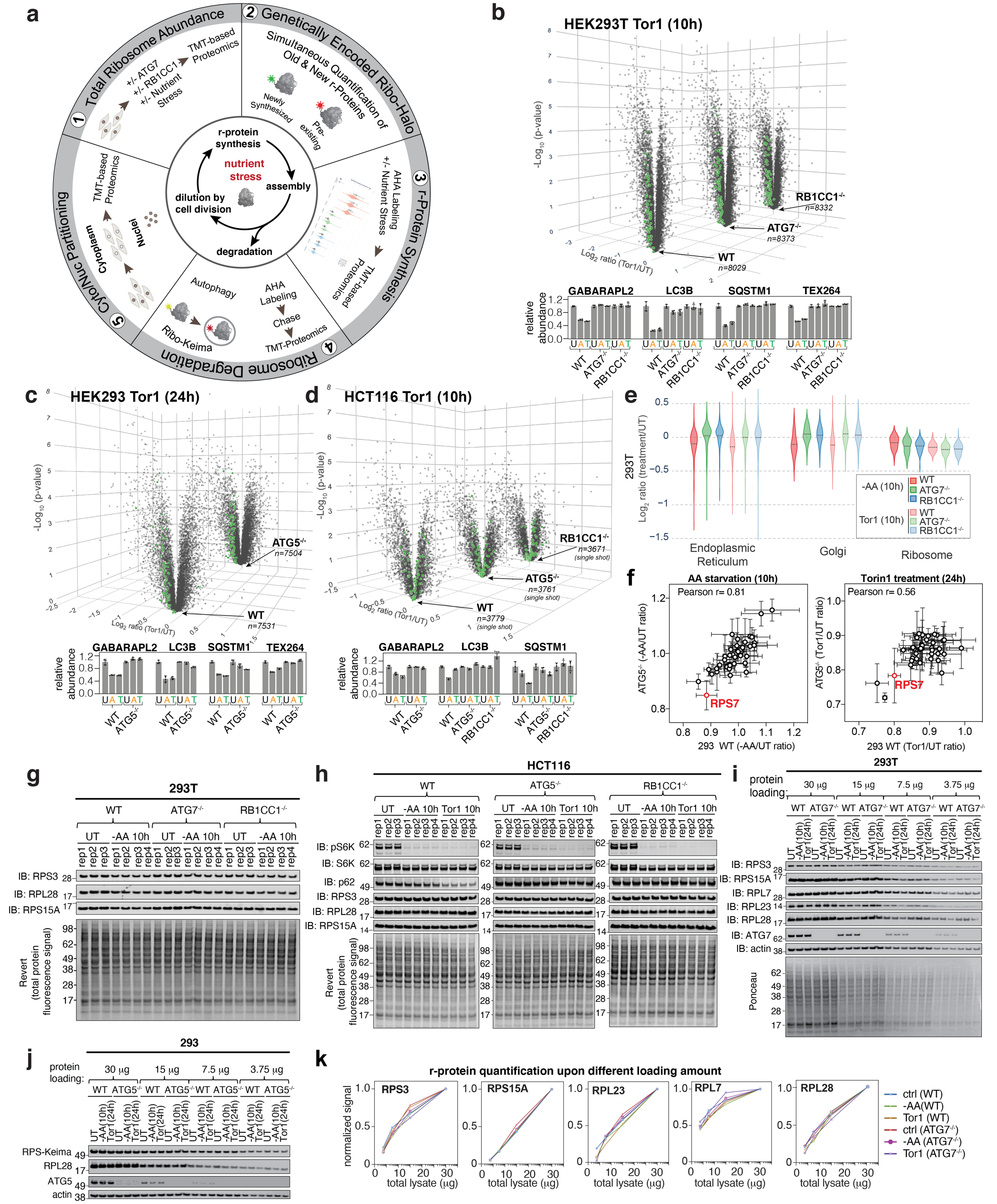
Competing interests J.W.H. is a founder and consultant for Caraway Therapeutics and a consultant for X-Chem, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2446-y.

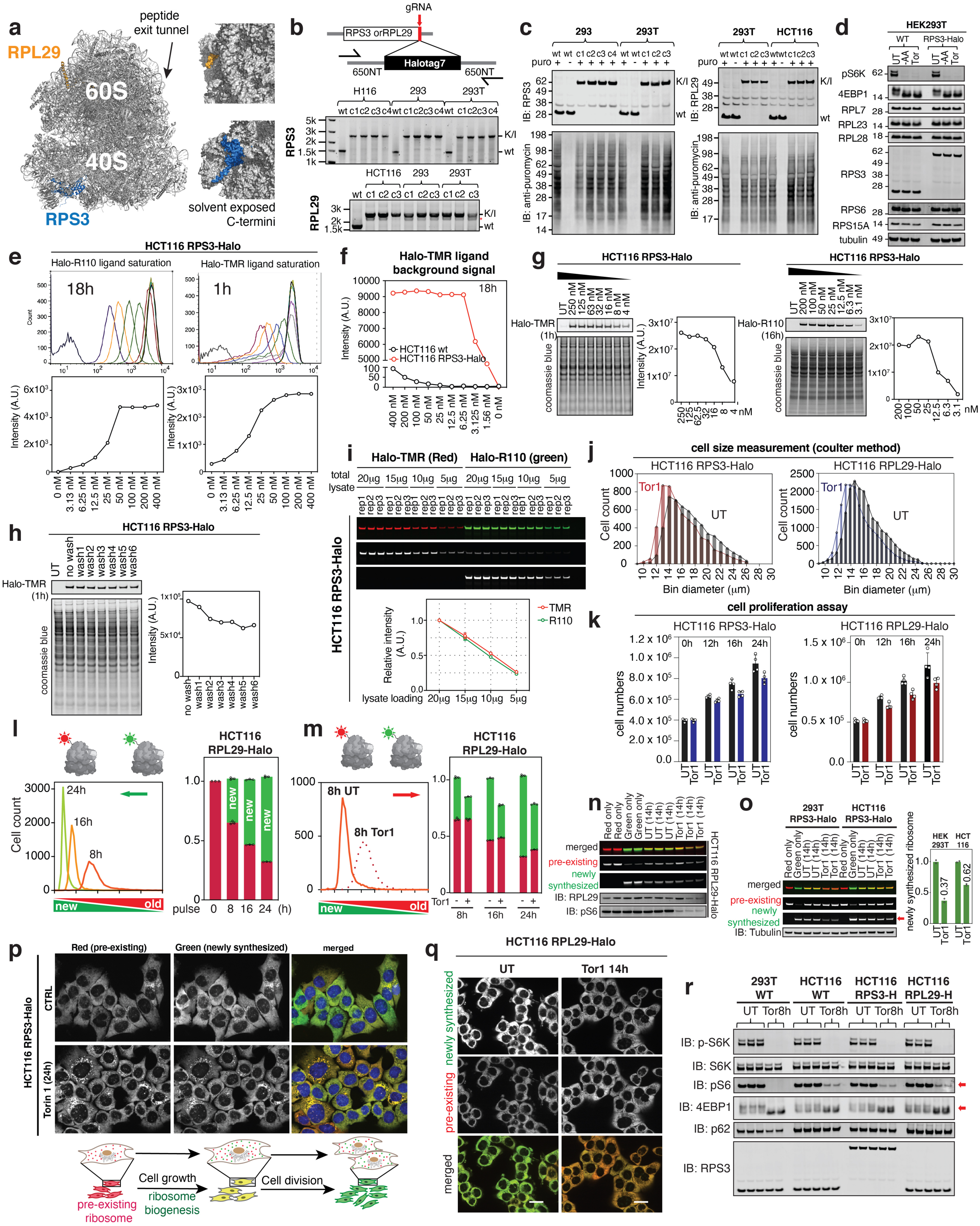
Correspondence and requests for materials should be addressed to J.W.H.

Peer review information *Nature* thanks Kris Gevaert, Suresh Subramani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



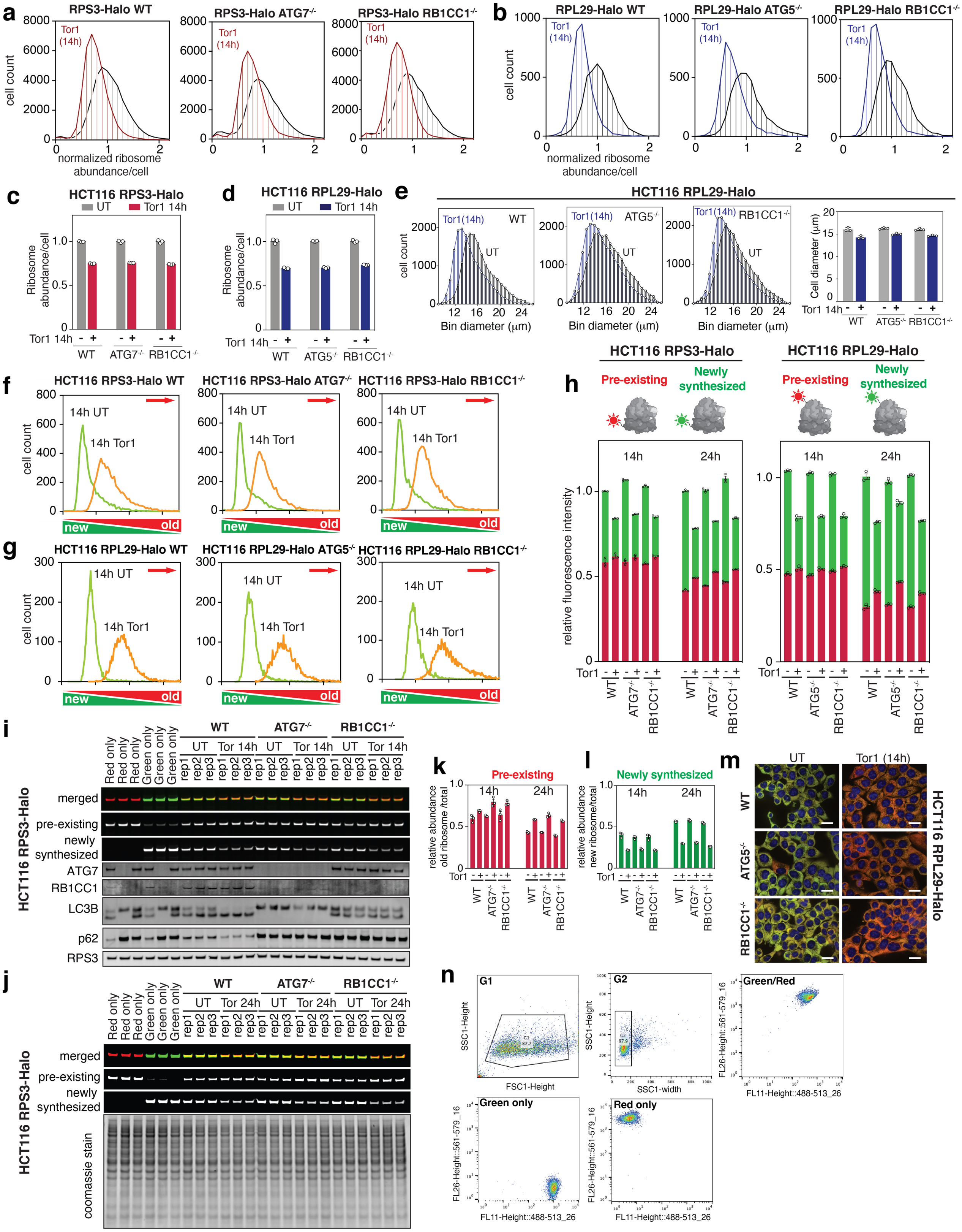
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Reduction of r-proteins during nutrient stress is not discernible by immunoblotting methods regardless of autophagy. a, Schematic of the ribosome analysis pipeline. **b**–**d**, (top panels) Volcano plots (-Log$_{10}$ p-value versus Log$_2$ ratio Tor1/UT) for 293T cells (n = 8029 proteins), *ATG7*$^{-/-}$ (n = 8373 proteins) or *RB1CC1*$^{-/-}$ (n = 8332 proteins) (**b**), HEK293 cells (n = 7531 proteins), *ATG5*$^{-/-}$ (n = 7504 proteins) (**c**), or HCT116 cells (n = 3779 proteins), *ATG5*$^{-/-}$ (n = 3761 proteins) or *RB1CC1*$^{-/-}$ (n = 3671 proteins) (**d**). n = 3 (UT); 4 (Tor1) biologically independent samples. P values were calculated by two-sided Welch's t-test (adjusted for multiple comparisons); for parameters, individual P values and q values, see Supplementary Table 1. Green dots represent r-proteins. Data for 293T cells are from ref. [12]. Histograms below the individual volcano plots show the mean ± s.d. of relative abundance of autophagy adaptors with or without nutrient deprivation. n = 3 (UT) or n = 4 (−AA and Tor1) biologically independent samples. U, untreated; A, −AA; T, Tor1. Data for 293T cells are from ref. [12]. **e**, The relative abundance changes[12] for proteins located in the ER, Golgi, or the ribosome in 293T cells treated as in **b** are plotted as a violin plot (n = 340, 349, 343, 340, 349, 343, 87, 89, 86, 87, 89, 86, 72, 75, 70, 72, 75, and 70 proteins, from left to right). R-protein abundance change is not affected by autophagy unlike other vesicular organelles. The violin curves represent the distribution and density of the indicated dataset (Centre-line: median; Limits: minima and maxima). **f**, Plots of relative abundance of individual r-protein in HEK293 cells upon either 10h of amino acid withdrawal (left) or Tor1 treatment (right). 39 r-proteins with less than ± 10% error range for every condition. Mean ± s.d. for n = 3 (UT) or n = 4 (−AA and Tor1) biologically independent samples. **g**, 293T cells with or without ATG7 or RB1CC1 were either left untreated, subjected to amino acid withdrawal (10h) or treated with Tor1 (10h) and whole cell extracts immunoblotted for the indicated proteins. **h**, HCT116 cells with or without ATG5 or RB1CC1 were treated as indicated, and whole cell extracts immunoblotted for the indicated proteins. **i**–**k**, Extracts from the indicated cells (30, 15, 7.5 or 3.75 μg) were immunoblotted with the indicated antibodies (**i**, **j**). The signal intensity for the indicated r-proteins as a function of quantity loaded was measured using Odyssey (**k**), showing no indication of signal saturation and no detectable difference between cells with or without active autophagy. Related to Fig. 1. The experiments shown in **g**–**j** were repeated more than three times independently and showed similar results. For gel source data, see Supplementary Fig. 1.



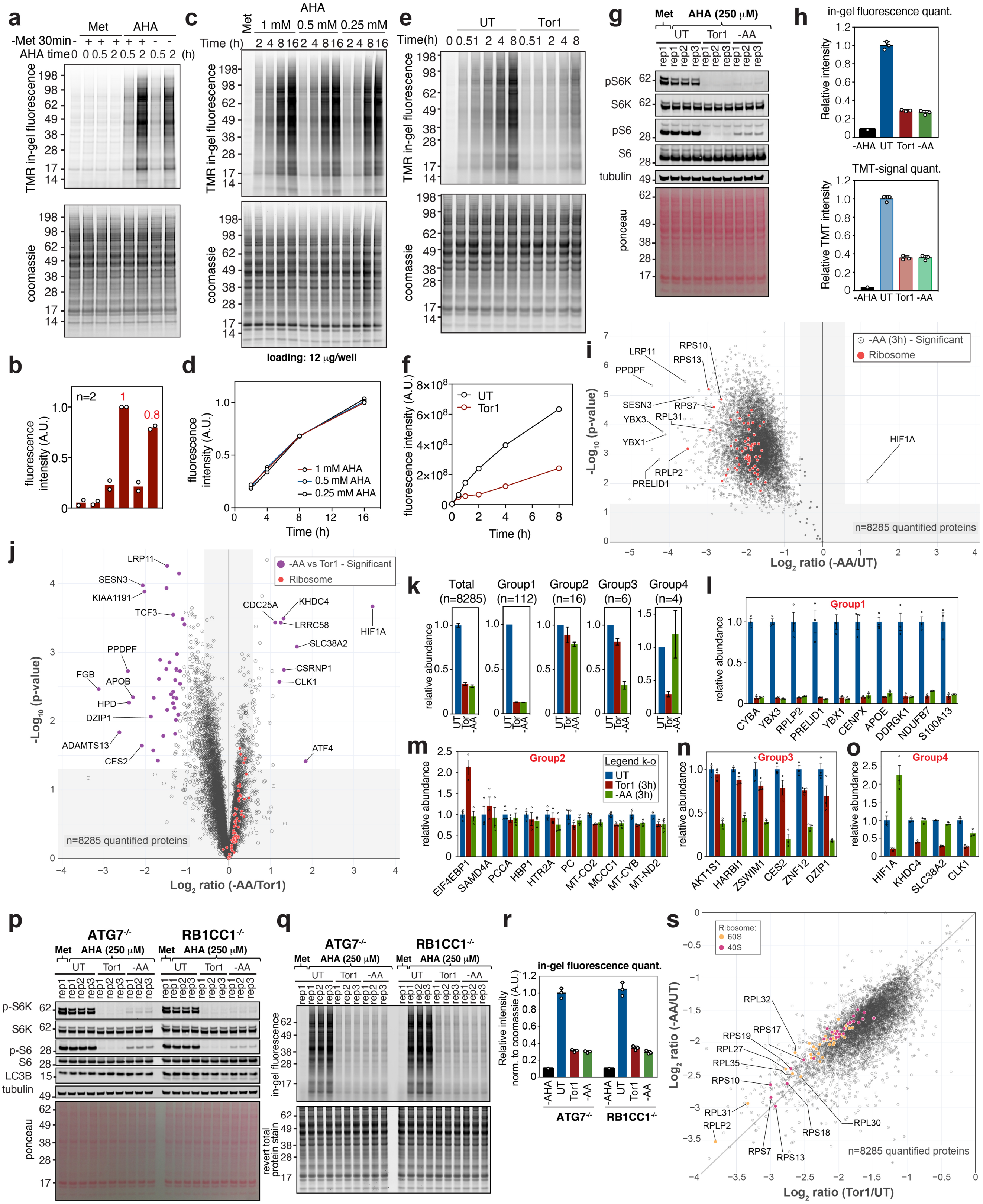
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Generation of Ribo–Halo reporters and extensive titration assays for quantification. a, The Halotag7 protein (referred to here as 'Halo') was endogenously tagged at the C terminus of RPL29 or RPS3 as the indicated r-proteins contain solvent-exposed C termini and are located far from the peptide exit tunnel based on the structure of an 80S complex. (PDB: 5AJ0). **b**, **c**, Gene editing of HCT116, HEK293 and 293T cells using CRISPR-Cas9 to fuse Halo with the C termini of RPS3 and RPL29. Homozygous incorporation of Halo was confirmed by genotyping (**b**). Extracts from the indicated cells were subjected to immunoblotting with RPS3 (**c**, left) or RPL29 (**c**, right). Protein translation efficiency of wild-type and Halo knock-in cells were compared using puromycin incorporation assay (**c**, bottom). **d**, Immunoblotting of WT or RPS3-Halo 293T cell lysates after nutrient stress, confirming no detectable difference between the two cell lines in response to mTOR inhibition. The full immunoblot is shown in Extended Data Fig. 9i. **e**–**g**, Halo-ligand titration assays with the indicated incubation time were performed using flow cytometry analysis (**e**, **f**) or in-gel fluorescence analysis (**g**) for the labelling saturation. In **f**, background signal from free Halo-ligand was measured using WT HCT116 cells in comparison to RPS3-Halo, confirming that the free ligand does not contribute to the observed fluorescence signal. **h**, HCT116 RPS3-Halo cells were incubated with 250 nM Halo-TMR ligand for 1h, washed for indicated numbers by incubating cells in ligand free medium for 20 min each time, followed by 17-h prolonged incubation in ligand free medium. **i**, Extracts from the RPS3-Halo HCT116 cells (20, 15, 10 or 5 μg) treated with the indicated Halo ligands were subjected to in-gel fluorescence analysis. The fluorescence signal intensity of each lane was directly proportional to the loading amount. We noted that R110 fluorophore was excited by epi-green excitation (520-545 nm) and detected in 577-613 nm. We subtracted this bleed-through signal for TMR quantification. **j**, Measurement of the effect of Tor1 on cell size with HCT116 RPS3-Halo (left) and RPL29-Halo (right) using Coulter Principle-based cell measurements. **k**, Cell proliferation assay. HCT116 RPS3-Halo and RPL29-Halo cells were grown in rich medium or Tor1 (200 nM) containing medium for 12, 16 and 24 h. The data estimates around 16-h cell division rate for untreated cells, and around 24-h cell division rate for Tor1 treated cells. Mean ± s.d. for $n = 4$ biologically independent experiments. **l**, Ratio of pre-existing to newly synthesized RPL29-Halo per cell plotted against cell populations as a frequency histogram (left). Average from the triplicate experiments plotted as a bar graph (right). Pre-existing Ribo–Halo proteins in HCT116 RPL29-Halo cells were labelled with TMR ligand (100 nM, 1 h), followed by the thorough washing and addition of 50 nM Green-ligand (also called R110-ligand). The newly synthesized RPL29-Halo was chased for 8, 16, and 24 h before flow cytometry analysis. Error bars represent s.d. See Methods for details. **m**, Pre-existing RPL29-Halo proteins were labelled with TMR ligand (100 nM, 1 h) in HCT116 cells, and the newly synthesized RPL29-Halo in the presence or absence of Tor1 (200 nM) were labelled with Green-ligand. The ratio of R110 to TMR signals plotted against cell populations (left), and the mean ± s.d. values from the triplicate experiments of 8, 16 and 24-h pulse chase plotted as a bar graph (right) are shown. **n**, In-gel fluorescence images of the cell extracts treated as in **m**. The same gels were then transferred to PVDF membranes for immunoblotting measurement of total RPL29 level. $n = 3$ biologically independent samples. **o**, In-gel fluorescence images of the cell extracts from 293T RPS3-Halo or HCT116 RPS3-Halo cells using the labelling strategy in **m**. Relative synthesis of RPS3-Halo with or without Tor1 is plotted on the right. Mean for $n = 2$ experiments. **p**, Live-cell imaging of HCT116 RPS3-Halo cells labelled with TMR (for pre-existing r-proteins) and Green (for newly synthesized r-proteins) ligands with or without Tor1 (200 nM, 24 h). Scale bar, 20 μm. **q**, Live-cell imaging of HCT116 RPL29-Halo cells labelled with TMR (for pre-existing r-proteins) and Green (for newly synthesized r-proteins) ligands with or without Tor1 (200 nM, 14h). Scale bar, 20 μm. **r**, The indicated cells were left untreated or incubated with Tor1 for 8 h before immunoblotting with the indicated antibodies. Related to Fig. 2. The experiments in **d**, **j** and **p**–**r** were repeated three times independently with similar results, and **b**, **c** and **e**–**i** were performed once. For gel source data, see Supplementary Fig. 1.



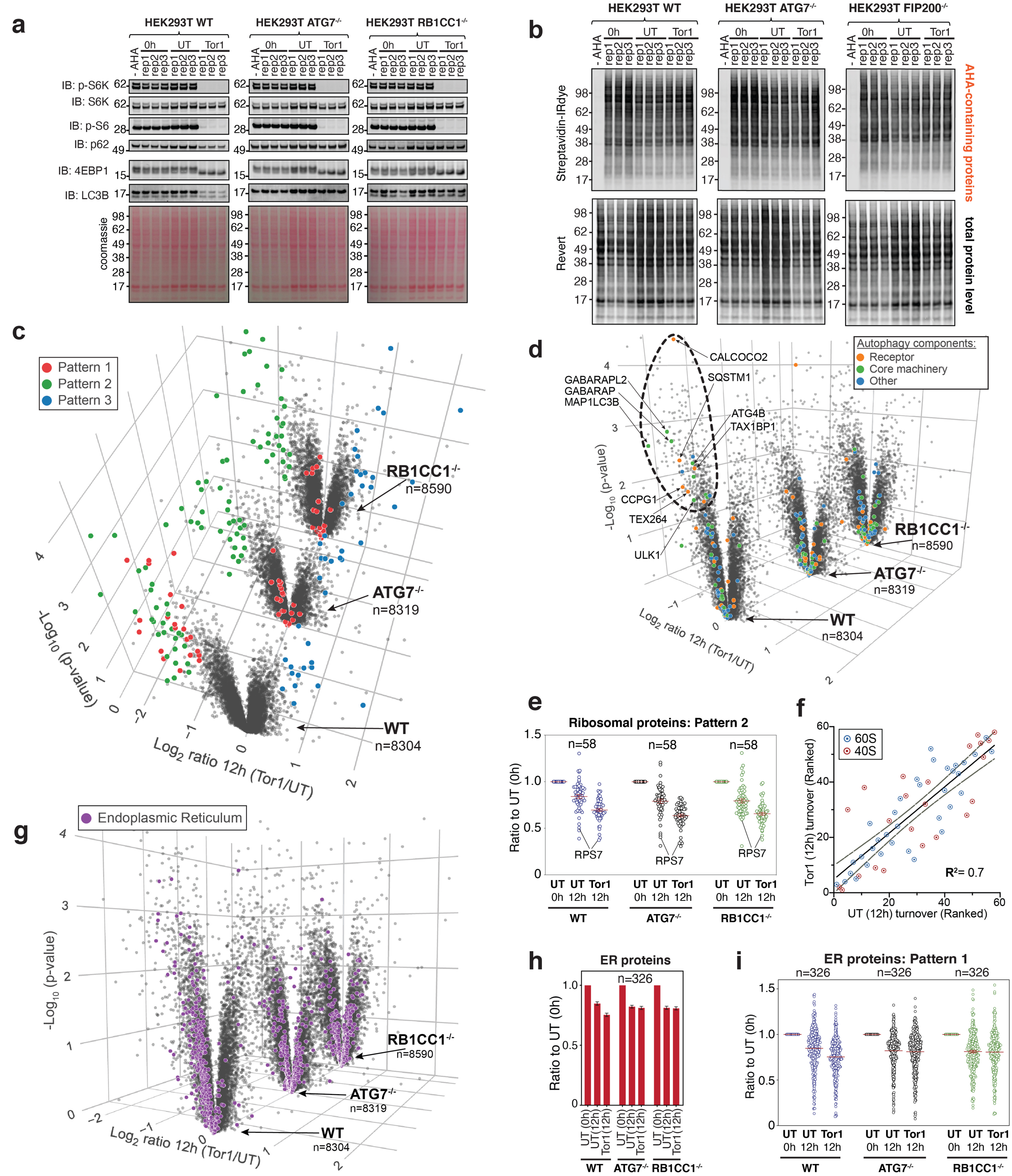
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Minimal contribution of ribophagy to control of r-protein synthesis and dilution by cell division in response to nutrient stress. **a**, **b**, Histogram of normalized TMR signal in RPS3-Halo (**a**) and RPL29-Halo (**b**) HCT116 cells with or without ATG8 conjugation (ATG7 for RPS3-Halo and ATG5 for RPL29-Halo) or RB1CC1 incubated with or without 200 nM Tor1 for 14 h, followed by 1 h TMR ligand treatment and flow cytometry analysis. >3x10^5 and >4x10^3 cells were analysed, respectively. **c**, **d**, Mean ± s.d. of the triplicate data from cells treated as in **a**, **b** are plotted, respectively. **e**, Effect of Tor1 treatment on cell size in HCT116 RPL29-Halo WT, *ATG5*$^{-/-}$ and *RB1CC1*$^{-/-}$ cells, as measured using Coulter Principle-based cell measurements. Mean ± s.d. of the triplicate data. **f**, **g**, Ratio of pre-existing (red) to newly synthesized (green) r-proteins per cell plotted against cell populations as a frequency histogram for RPS3-Halo (**f**) or RPL29-Halo (**g**) HCT116 cells with or without ATG5, ATG7 or RB1CC1 based on the labelling scheme in Fig. 2f. **h**, Quantification of relative amounts of pre-existing and newly synthesized r-proteins from data in **f**, **g**. Mean ± s.d., $n = 3$ biologically independent experiments. **i**, **j**, HCT116 RPS3-Halo cells with or without ATG7 or RB1CC1 were left untreated or treated with Tor1 for 14h (**i**) and 24h (**j**) using the Halo tagging scheme in Fig. 2f. Extracts were subjected to SDS–PAGE and in-gel fluorescence analysis, followed by immunoblotting with the indicated antibodies. **k**, **l**, Quantification of relative amounts of pre-existing and newly synthesized r-proteins from data in **i**, **j**. Mean ± s.d., $n = 3$ biologically independent experiments. **m**, Live-cell imaging of HCT116 RPL29-Halo cells with indicated genotypes labelled with TMR (for pre-existing r-proteins) and Green (for newly synthesized r-proteins) ligands with or without Tor1 (200 nM, 14h). Scale bar, 20 µm. **n**, An example of gating strategy used for flow cytometry analysis. Green and Red only control experiments are shown at the bottom. Experiments in **m**, **n** were repeated more than three times independently with similar results. For gel source data, see Supplementary Fig. 1.



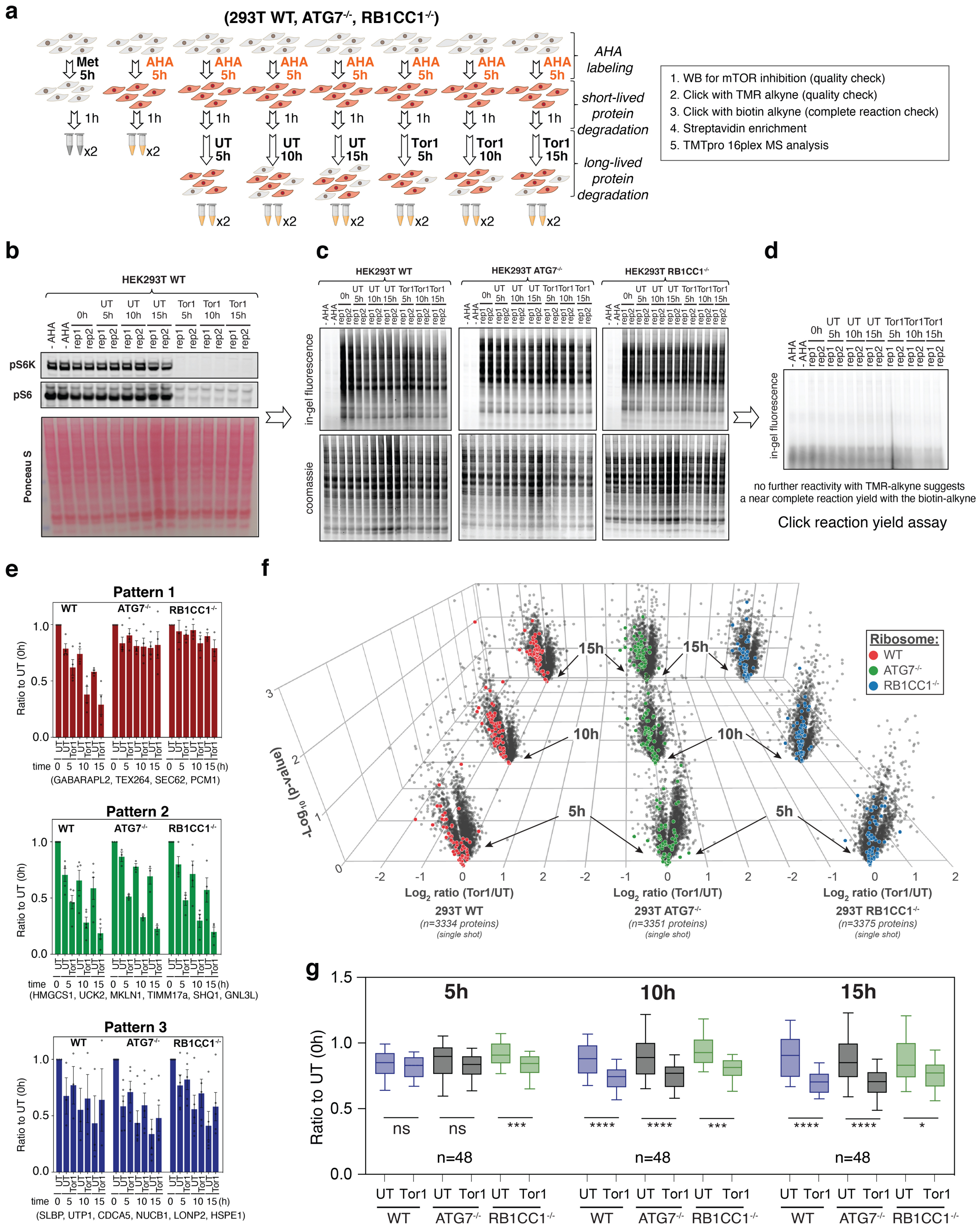
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Global decoding of protein translation during nutrient stress via independent AHA-TMT methods. a, **b**, Total AHA incorporation levels with or without prior Methionine starvation (30 min) were compared using 293T cells grown in Met or AHA (250 µM) for the indicated duration, followed by click with TMR alkyne and in-gel fluorescence analysis (**a**). The quantification of duplicate experiments is shown in **b**. **c**, **d**, 293T cells were grown in medium with Met (250 µM) or the indicated concentration of AHA for the indicated time periods. Extracts were clicked with TMR and subjected to SDS–PAGE before in-gel TMR fluorescence analysis (**c**). TMR intensity was quantified in **d**. **e**, **f**, 293T cells grown in AHA (250 µM) with or without To1 for the indicated time periods and extracts clicked with TMR before processing as in **c**. The effect of Tor1 on TMR fluorescence is quantified in **f**. **g**, 293T cells were incubated with or without amino acid withdrawal or Tor1-containing medium in the presence of Met or AHA (250 µM each), as in Fig. 3c. Cell extracts were subjected to SDS–PAGE followed by immunoblotting. Three biologically independent samples are shown in the same blot. **h**, TMR signals in Fig. 3c. were quantified as described in Methods and plotted in top panel, and the relative TMT signal of the total biotinylated proteome in Fig. 3d. is plotted in bottom panel. Centre data are mean ± s.d. *n* = 1, 3, 3 and 3 biologically independent samples, from left to right for top and bottom panels. **i**, Translatome analysis. 293T cells were carried through the workflow in Fig. 3b with amino acid withdrawal and extracts clicked with biotin before enrichment on streptavidin and TMT-based proteomics. Plot of -Log$_{10}$ p-value versus Log$_2$ ratio −AA/untreated is shown for *n* = 8285 proteins. *n* = 3 (UT; −AA) biologically independent samples. **j**, Translatome plot of -Log$_{10}$ p-value versus Log$_2$ ratio −AA/Tor1 is shown (*n* = 8285 proteins). r-proteins skewed to the right side of the volcano plot indicates that Tor1 suppresses the translation of r-proteins more strongly than amino acid withdrawal, unlike the majority of the proteome. *n* = 3 (−AA; Tor1) biologically independent samples. *P* values in **i** and **j** were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 2. **k**–**o**, The relative abundance of all quantified biotinylated proteins are shown in **k**. Proteins in individual groups are shown in **l**–**o**. **l**: representative proteins showing over twofold more reduction in translation than the average translation in both Tor1 and −AA conditions, **m**: proteins with over twofold less reduction in translation than the average translation in both Tor1 and −AA conditions, **n**: proteins showing twofold less reduction only in Tor1 condition, with *p*-value (Supplementary Table 2) between Tor1 and −AA was less than 0.05. **o**: proteins showing twofold less reduction only in −AA condition, with *p*-value (Supplementary Table 2) between Tor1 and −AA was less than 0.05. *n* = 3 biologically independent samples per conditions (UT; −AA; Tor1). Centre data are mean ± S.E.M. **p**–**r**, 293T cells lacking either *ATG7* or *RB1CC1* were subjected to amino acid withdrawal or Tor1 treatment for 3 h in the presence of 250 µM Met or AHA. Cell extracts were subjected to SDS–PAGE followed by immunoblotting with the indicated antibodies (**p**) or clicked with TMR and in-gel fluorescence analysis (**q**). TMR signals were quantified as described in Methods (**r**). Centre data are mean ± s.d. *n* = 1, 3, 3 and 3 biologically independent samples, from left to right. **s**, Correlation plot (Log$_2$ ratio of treated/untreated) for the translatome upon either Tor1 treatment or amino acid withdrawal. Related to Fig. 3. Experiments in **c**–**f** were performed once and **p** was performed three times independently with similar results. For gel source data, see Supplementary Fig. 1.



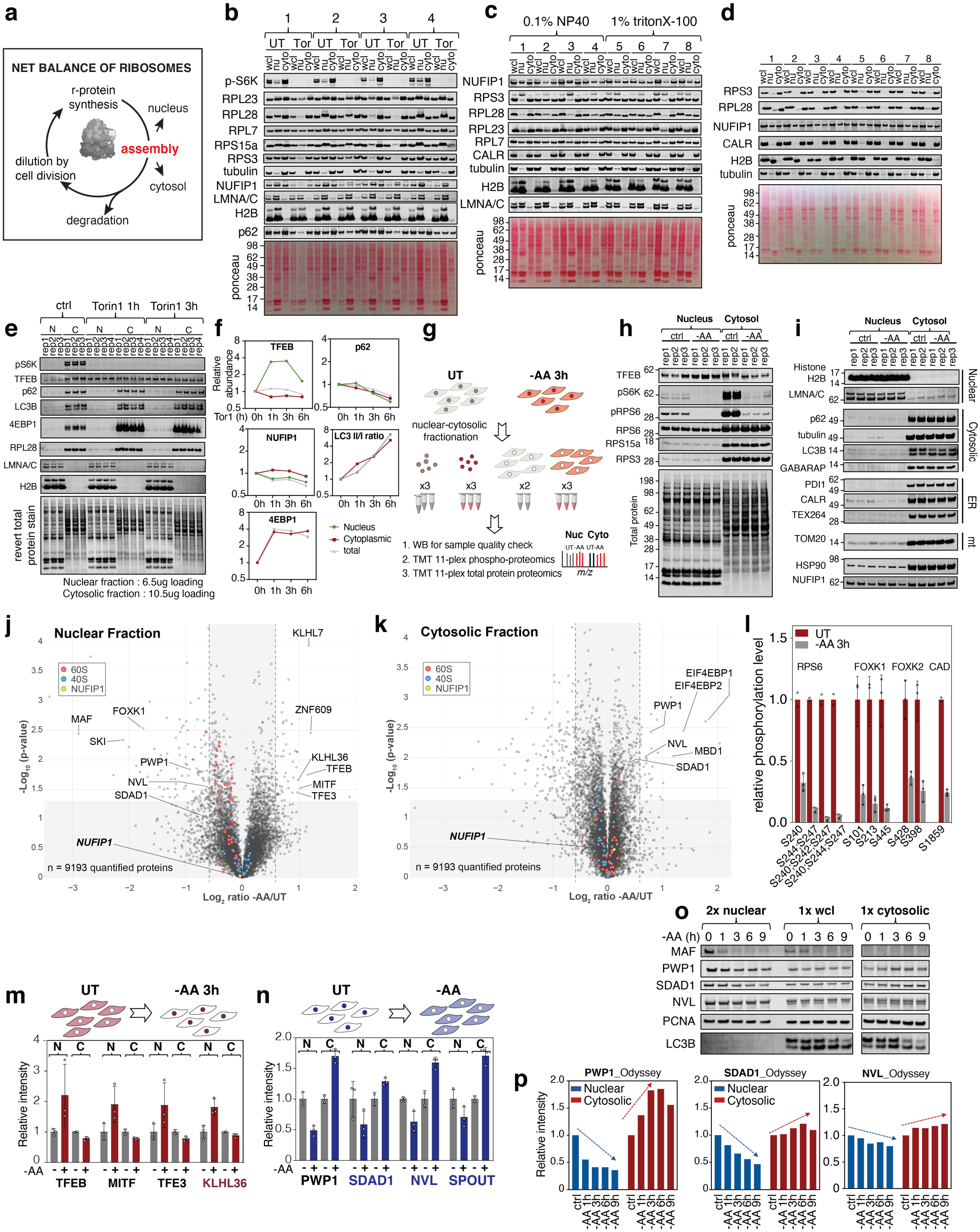
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Global decoding of protein degradation during nutrient stress via independent AHA-TMT methods. a, Extracts from cells as described in Fig. 3g were subjected to immunoblotting with the indicated antibodies to demonstrate suppression of the mTOR activity by Tor1. **b**, Biotinylated extracts as described in Fig. 3g were subjected to immunoblotting with a fluorescent streptavidin conjugant, showing the pre-existing proteome (Streptavidin-IRdye) versus the total proteome (Revert total protein stain) in the lysates. *n* = 3 biologically independent samples are shown in **a** and **b**. **c**, Volcano plots (-Log_{10} p-value versus Log_2 Tor1/UT 12h) for 293T WT (*n* = 8304 proteins), *ATG7*$^{-/-}$ (*n* = 8319 proteins), and *RB1CC1*$^{-/-}$ (*n* = 8590 proteins) cells as in Fig. 3i, but with pattern 1 proteins in Fig. 3h coloured as red dots, pattern 2 proteins coloured as green dots, and pattern 3 proteins coloured as blue dots. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 3. *n* = 3 (UT; Tor1) biologically independent samples. **d**, Volcano plots as in **c**, but with autophagy related proteins coloured as indicated. Autophagy adaptors dramatically degraded upon Tor1 treatment only in WT cells are shown in the circle. **e**, Plots of individual ratio for AHA-labelled r-proteins using the protocol in Fig. 3g in WT, *ATG7*$^{-/-}$, or *RB1CC1*$^{-/-}$ 293T cells. r-proteins with STDEV <0.3 for every condition are selected (*n* = 58 r-proteins). Centre data are mean ± s.d. from 3 biologically independent samples for each condition. **f**, Relative turnover rates for individual r-proteins from **e**, with or without mTOR inhibition have a correlation of R^2~0.7 (*n* = 58 r-proteins). Grey dotted lines are 95% confidence intervals of the best-fit line (solid black line) result from a simple linear regression analysis. **g**, Volcano plots as in **d**, but with ER resident proteins coloured in purple. **h**, **i**, Individual ratio for *n* = 326 AHA-labelled ER membrane resident proteins using the protocol in Fig. 3g in WT, *ATG7*$^{-/-}$, or *RB1CC1*$^{-/-}$ 293T cells (**h**) and plots for all individual ER resident proteins data points used in h (**i**). Centre data are mean ± s.e.m. Related to Fig. 3. For gel source data, see Supplementary Fig. 1.



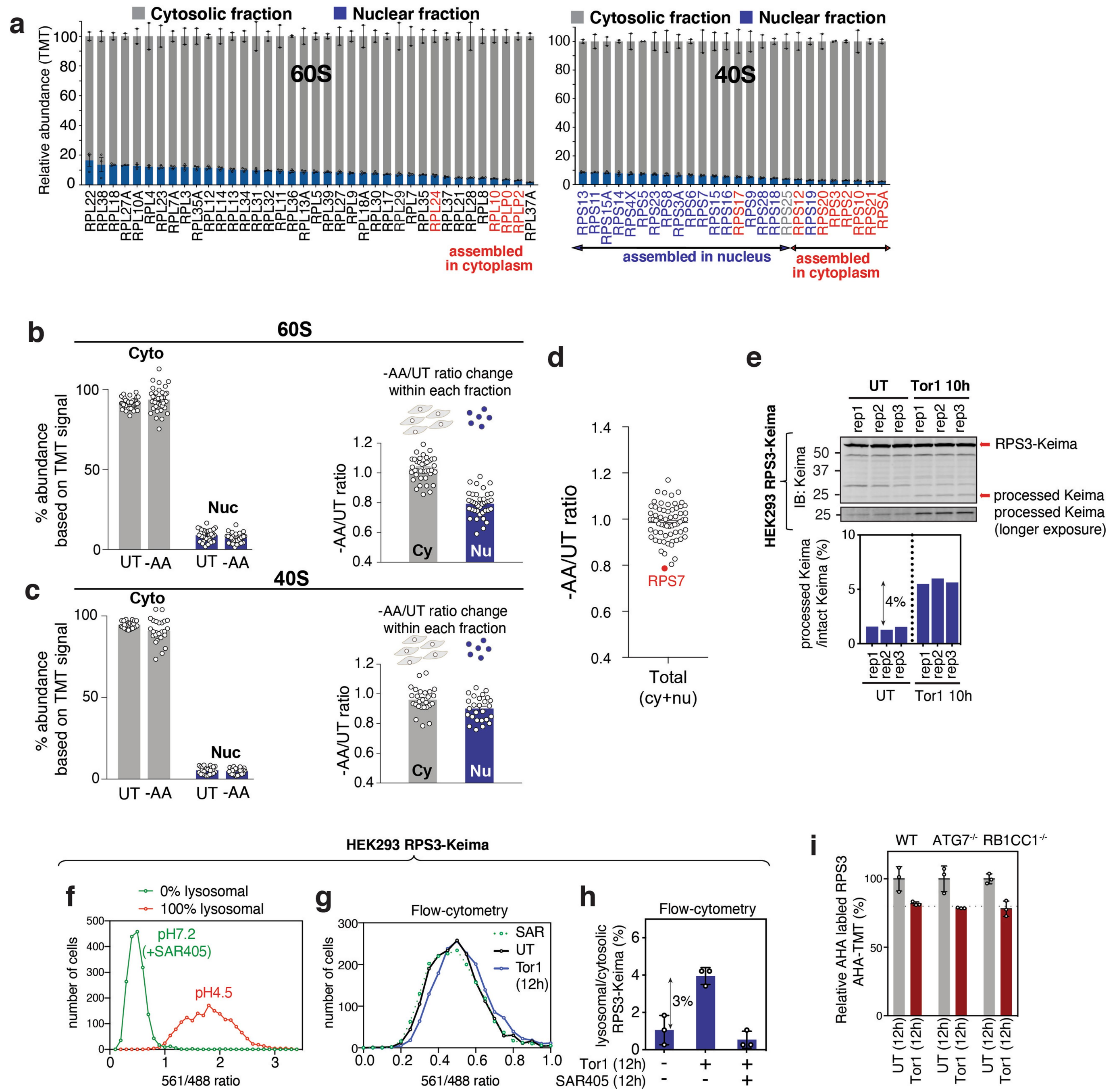
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Protein degradation time-course experiment during nutrient stress via AHA-TMT methods. a, Schematic of AHA-based degradomics time-course to examine r-protein turnover during nutrient stress with or without functional autophagy. **b**, Lysates from 293T cells treated as in **a**, were subjected to SDS–PAGE followed by immunoblotting to confirm proper mTOR inhibition. **c**, Cell extracts were reacted with TMR alkyne for click reaction followed by in-gel fluorescence analysis. **d**, Click reaction yield across the replicates was confirmed indirectly. In brief, cell extracts treated as in **a**, were clicked with biotin alkyne followed by streptavidin capture. The proteins in the flow-through was precipitated, resuspended in 2% SDS, then clicked with TMR-alkyne. **e**, Patterns of protein turnover, as described in the main Fig. 3h, in the time-course experiment. mean ± s.e.m. Proteins analysed (*n* = 4 top; *n* = 6 middle and bottom) are shown below. **f**, Volcano plots for the indicated time (-Log$_{10}$ p-value versus Log$_2$ Tor1/UT) in 293T WT (*n* = 3334 proteins), *ATG7*$^{-/-}$ (*n* = 3351 proteins), and *RB1CC1*$^{-/-}$ (*n* = 3375 proteins) cells. r-proteins, red, green or blue. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 3. **g**, A box plot of the individual ratio for AHA-labelled r-proteins using the protocol in **a** (centre line, median; box limits correspond to the first and third quartiles; whiskers, 10-90 percentiles range). *n* = 48 r-proteins that were quantified across WT, *ATG7*$^{-/-}$ and *RB1CC1*$^{-/-}$ 293T cells. two-sided *t*-test, *P* = 0.1514, 0.2818, 0.0005, 0.000016, 0.000051, 0.0005, 0.000001, 0.000020, 0.0105 from left to right; NS: non-significant, **P* < 0.1, ****P* < 0.001, *****P* < 0.0001. Experiments in **b**–**d** were replicated twice independently and showed similar results. For gel source data, see Supplementary Fig. 1.

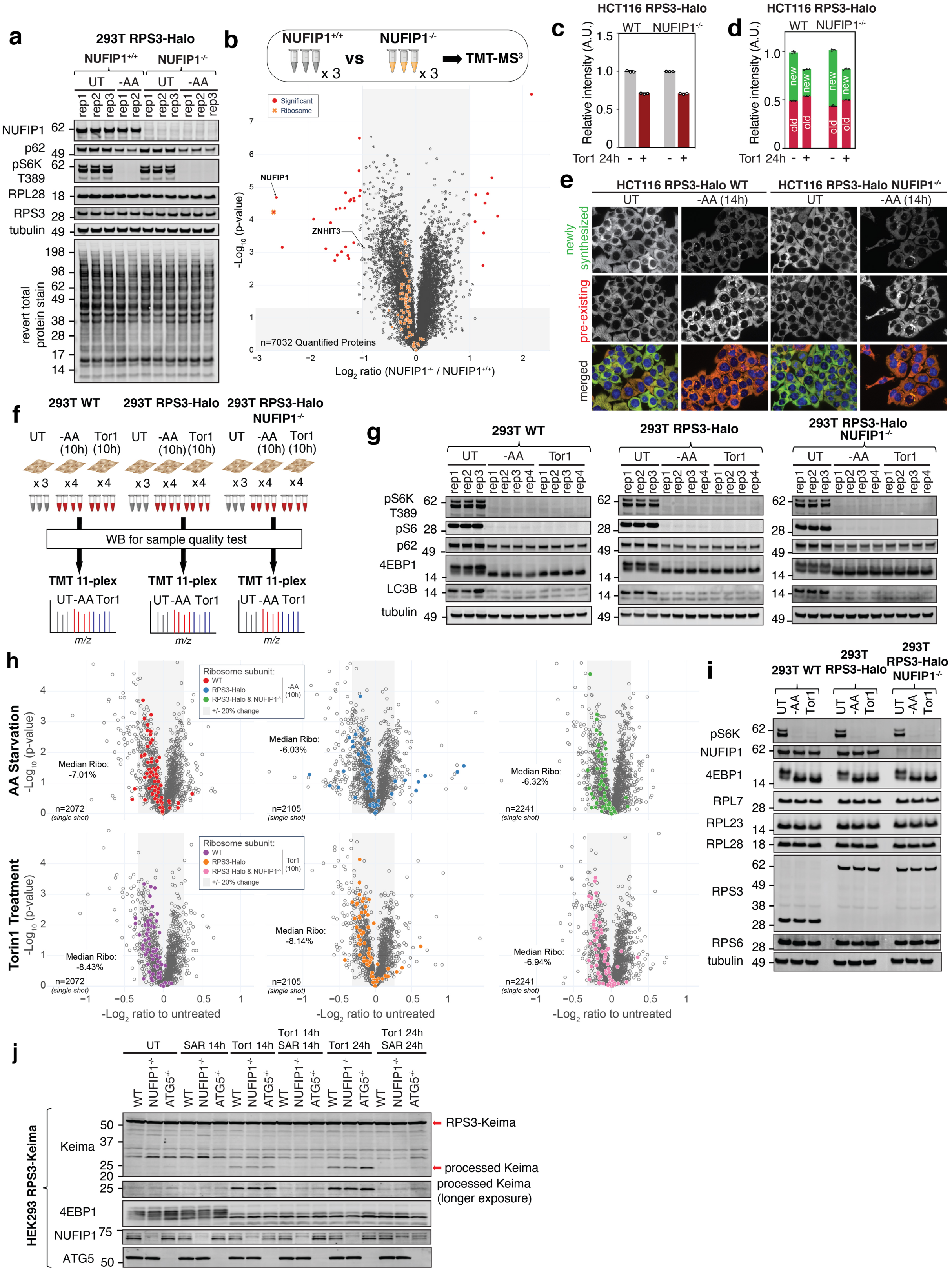


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Optimization of nuclear–cytosolic partitioning during nutrient stress. a, Contribution of cytoplasmic and nuclear partitioning to net ribosome balance upon nutrient stress. **b**, **c**, Optimization of the nuclear and cytosolic fraction partitioning method using 293T cells. Side by side comparison of four previously published methods was performed. Surfactant used in method1: 1% Triton, method 2: 0.1% Triton, method 3: 0.1% NP40, method 4: 0.05% NP40. For more details on methods 1-4, see Methods. Methods 2 and 3 were further compared in **c**. 1: ctrl, 2: Leptomycin B (20 nM, 16h), 3: $MgCl_2$ added in lysis buffer, 4: 2+3, 5: ctrl, 6: Leptomycin B (20 nM, 16h), 7: $MgCl_2$ added in lysis buffer, 8: additional pipetting. **d**, Effect of the centrifugal velocity and duration on nuclear-cytosol partitioning is shown. 1: 13K rpm, 10sec, 2: 10K rpm, 10sec, 3: 7K rpm, 10sec, 4: 7K rpm, 30sec, 5: 5K rpm, 60sec, 6: 5K rpm, 180sec, 7: 3K rpm, 60sec, 8: 3K rpm, 180sec. **e**, **f**, Lysates collected after Tor1 treatment for 0, 1 or 3 h subjected to the optimized nuclear-cytosol partitioning, followed by immunoblotting against the indicated antibodies (**e**). Quantification measured by Odyssey shown in **f**. **g**, Scheme depicting strategy for quantitative analysis of changes in nuclear and cytosolic protein abundance in 293T cells in response to short period (3 h) of amino acid withdrawal. **h**, **i**, Biochemical characterization of nuclear and cytosolic 293T cell fractions in response to amino acid withdrawal. Extracts (15 µg of cytosol and nuclei) were separated by SDS–PAGE and immunoblots probed with the indicated antibodies (see Methods). **j**, **k**, Volcano plots (-Log_{10} p-value versus Log_2 −AA/UT) for nuclear (**j**) or cytosolic (**k**) proteins ($n = 9193$ proteins) quantified by TMT-based proteomics. P values were calculated by two-sided Welch's t-test (adjusted for multiple comparisons); for parameters, individual P values and q values, see Supplementary Table 4. Nuclear fraction (**j**) $n = 3$ (UT; −AA), cytosolic fraction (**k**) $n = 2$ (UT); $n = 3$ (−AA) biologically independent samples. **l**, Relative RPS6, FOXK1, FOXK2, and CAD phospho-peptides abundance quantified by TMT-based proteomics confirming strong inhibition of mTOR by Tor1. Centre data are mean ± s.d. $n = 2$ for UT RPS6 and CAD, and 3 for the rest. **m**, **n**, Relative abundance of proteins that translocate either from cytosol to nucleus (**m**) or from nucleus to cytosol (**n**) after 3h amino acid withdrawal, including proteins linked with nutrient dependent transcription (TFEB, MITF, TFE3 – accumulating in the nucleus), and proteins involved in ribosome assembly (PWP1, SDAD1, NVL – exported from the nucleus to the cytosol). Centre data are mean ± s.d. $n = 3, 2, 3$ and 3 biologically independent samples, from left to right for each indicated proteins. **o**, **p**, 293T cells were treated with −AA medium for the indicated time period, partitioned into nuclear and cytosolic fractions, followed by immunoblotting against the indicated antibodies (**o**). Odyssey quantification shown in **p**. Experiments shown in **b**–**d** were performed once, **e**, **h**, **i** were performed more than three times independently with similar results, and **o** was performed twice with similar results. For gel source data, see Supplementary Fig. 1.



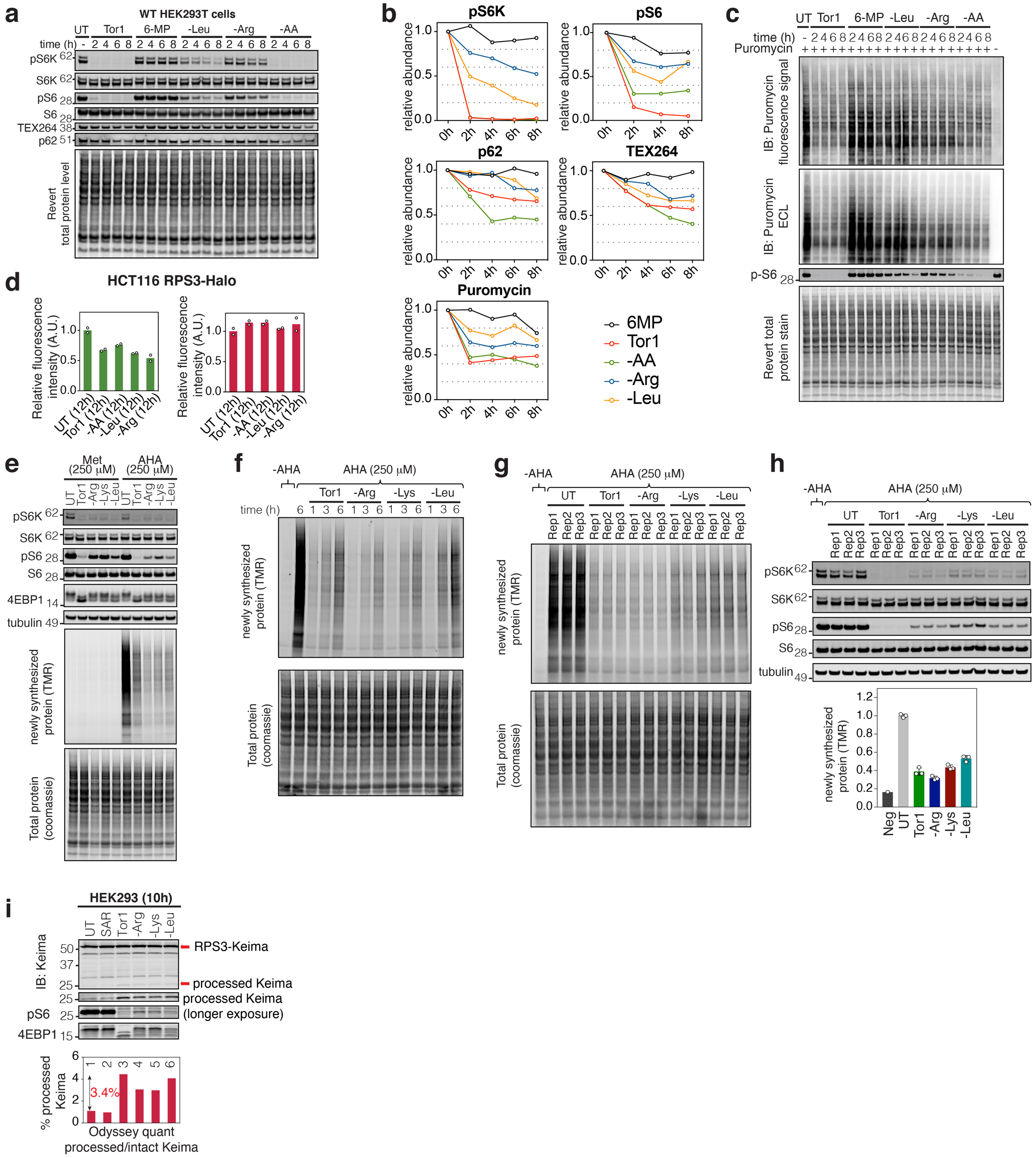
Extended Data Fig. 8 | Contribution of nuclear–cytosolic partitioning to r-protein abundance during nutrient stress and ribophagy flux measurement using the Ribo–Keima system. a, Relative abundance of individual r-proteins from the 60S subunit (left) or 40S subunit (right) with the cytosol fraction in grey and the nuclear fraction in blue. (less than ± 10% error range for every replicate) Individual r-proteins that are thought to assemble onto the ribosome either late in the assembly process or specifically in the cytosol are indicated in red font. mean ± s.d. $n = 2$ for cytosolic fraction and 3 for nuclear fraction as shown in Extended Data Fig. 7g. **b**, **c**, Abundance of nuclear and cytosolic r-proteins after amino acid withdrawal (3 h). 60S subunits are on top ($n = 38$), and 40S subunits are at the bottom ($n = 26$). Right panels indicate the relative r-protein abundance change normalized by UT of cytosolic or nuclear fraction. mean ± s.e.m. **d**, −AA/UT ratio of individual r-proteins from both nuclear and cytosolic fractions collected after 3 h amino acid starvation indicates heterogenous distribution with RPS7 most strongly down regulated. $n = 64$, mean ± s.e.m. **e**, Lysates from HEK293 RPS3-Keima cells after the indicated nutrient stress were immunoblotted against anti-Keima antibody (top). Abundance ratio of the processed Keima to the intact Keima measured by Odyssey is plotted (bottom). **f**, Flow cytometry analysis of HEK293 RPS3-Keima cells to obtain normalization factors. To achieve a condition in which cells have 0% of the ribosomes in the lysosome, RPS3-Keima cells were treated with SAR405 for 10 h and collected in pH7.2 FACS buffer. To achieve a theoretical condition in which 100% ribosomes are present in the lysosome, the cells were incubated in pH4.5 FACS buffer containing 0.1% Triton-X. We used the 561/488 ratio from these two measurements to calculate the % lysosomal ribosomes in **g** and **h**. $n = 1742$ cells for each. **g**, **h**, HEK293 RPS3-Keima cells were left untreated or treated with Tor1 (200 nM) in the presence or absence of SAR405 (1 µM) for 12 h. 561 nm ex to 488 nm ex Keima signal was measured by flow cytometry and plotted as either a frequency histogram (**g**, $n = 1742$ cells for each) or a bar graph (**h**, $n = 3$ biologically independent samples, mean ± s.d.). **i**, TMT-based quantification of endogenous RPS3 abundance in 293T WT, *ATG7*$^{-/-}$, and *RB1CC1*$^{-/-}$ cells treated and processed as in Fig. 3g. $n = 3$ biologically independent samples. mean ± s.d. Experiments in **e**, **g**, **h** were repeated three times independently with similar results, and **f** was repeated once. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 9 | See next page for caption.

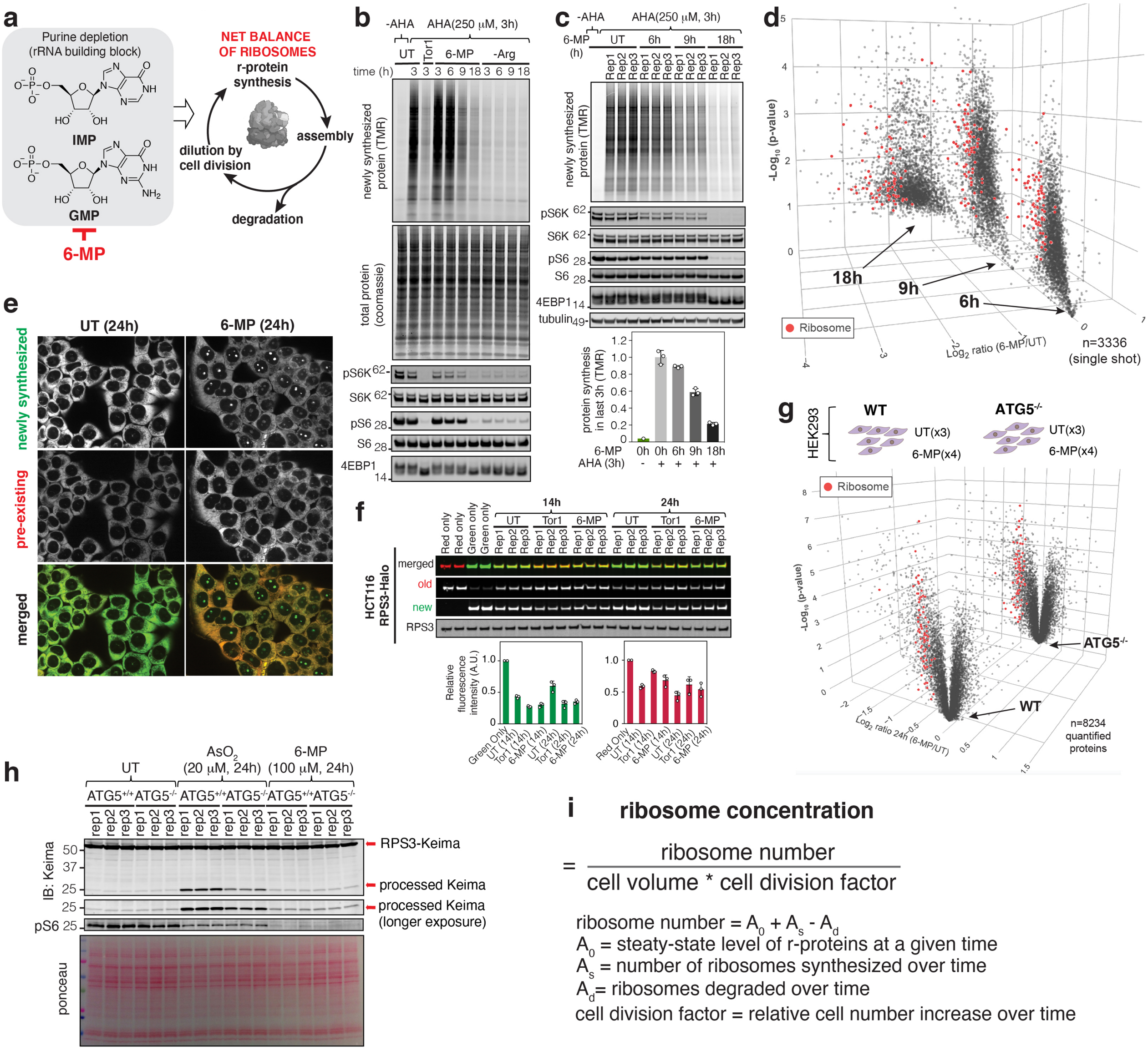
Extended Data Fig. 9 | NUFIP1 deletion does not affect the ribosome inventory changes with nutrient stress. a, Extracts from the 293T cells with or without NUFIP1 immunoblotted with the indicated antibodies after 10h of amino acid withdrawal, showing the same level of mTOR inhibition and r-proteins abundance regardless of the *NUFIP1* deletion. Three biologically independent samples are blotted except two samples for *NUFIP*$^{+/+}$ −AA condition. **b**, Volcano plot (-Log$_{10}$ p-value versus Log$_2$ ratio (*NUFIP1*$^{-/-}$/WT)) of 293T cells with or without *NUFIP1* deletion (*n* = 7032 proteins). *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 5. *n* = 3 biologically independent samples per genotype. **c**, Normalized TMR signal in HCT116 RPS3-Halo *NUFIP1*$^{+/+}$ or $^{-/-}$ cells incubated with or without 200 nM Torin for 24 h, followed by 1 h TMR ligand treatment and flow cytometry analysis. Mean ± s.d. of the triplicate data are plotted. **d**, Average ratio of pre-existing to newly synthesized RPS3-Halo per cell with or without NUFIP1 plotted as a bar graph. Pre-existing Ribo–Halo proteins were labelled with TMR ligand (100 nM, 1 h), followed by the addition of 50 nM Green-ligand. *n* = 3 biologically independent samples. mean ± s.d. **e**, Live-cell imaging of indicated Ribo–Halo cells with or without NUFIP1 labelled with TMR (for pre-existing r-proteins) and Green (for newly synthesized r-proteins) ligands with or without Tor1 (200 nM, 14 h). Scale bar, 20 µm. **f**, Schematic description of the triple TMT-MS analysis of the whole cell lysates gathered from WT 293T or RPS3-Halo 293T cells with or without NUFIP1 after nutrient stress for 10 h. **g**, Lysates of cells treated as in **f** were immunoblotted against the indicated antibodies for quality control, showing that mTOR activity was properly inhibited in all three cell types. **h**, Volcano plots (-Log$_{10}$ p-value versus Log$_2$ ratio Nutrient stress/Untreated) of the cells treated as in **f** (WT *n* = 2072 proteins; RPS3-Halo *n* = 2105 proteins; RPS3-Halo and NUFIP1$^{-/-}$ *n* = 2241 proteins). Introducing HaloTag at the endogenous locus did not alter the mTOR inhibition nor ribosome abundance change after nutrient stress. Deletion of NUFIP1 did not show detectable difference either. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 5. *n* = 3 (UT); 4 (−AA, Tor1) biologically independent samples per cell line. **i**, Immunoblotting of the cell lysates prepared as in **f** shows that introducing HaloTag at the endogenous locus did not alter the mTOR inhibition nor ribosome abundance change after nutrient stress. Deletion of NUFIP1 did not show detectable difference either, consistent with the TMT-MS analysis. **j**, Keima processing assay using the lysates from HEK293 RPS3-Keima WT or *NUFIP*$^{-/-}$ cells after the indicated nutrient stress. Also see Supplementary Table 5. Experiments in **e** were repeated twice with similar results, and **g**, **i**, **j** were repeated three times independently with similar results. For gel source data, see Supplementary Fig. 1.

WT HEK293T cells were used in all panels, otherwise indicated.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Systematic analysis of r-protein homeostasis in response to withdrawal of single amino acids. **a**, 293T cells were treated with either Tor1 or 6-MP or were incubated in medium lacking Leu, Arg, or amino acids for the indicated times and cell extracts subjected to immunoblotting with the indicated antibodies. **b**, Quantification of the WB data in **a** and **c**. **c**, Puromycin incorporation assay after treating 293T cells with the indicated medium and time course. Immunoblotting against anti-puromycin antibody was probed using either infrared fluorophore labelled second antibody coupled with Odyssey or Horseradish peroxidase labelled second antibody coupled with ECL for comparison. **d**, Quantification of the relative abundance of pre-existing (red) and newly synthesized (green) RPS3-Halo in Fig. 4b. mean, *n* = 2. **e**, Cell extracted treated as indicated were analysed for either mTOR inhibition or translatome using AHA incorporation assay coupled with TMR-alkyne click. **f**, Time-course protein synthesis assay was performed using AHA clicked with TMR-alkyne method after the indicated nutrient stress. **g**, **h**, Cells treated as in Fig. 4c were clicked with TMR-alkyne and analysed by in-gel fluorescence signal (**g**). Immunoblot assays using the indicated antibodies for quality control is shown in **h** (top), and relative TMR signal is plotted below. Centre data are mean ± s.d. *n* = 1, 3, 3, 3, 3 and 3 biologically independent samples, from left to right. **i**, Lysates from HEK293 RPS3-Keima cells after the indicated nutrient stress were immunoblotted against antibodies for Keima or mTOR substrates (top). Abundance ratio of the processed Keima to the intact Keima is plotted (bottom). Related to Fig. 4. Experiments in **a**, **c**, **g**, **i** were repeated three times independently with similar results, and **e**, **f**, **i** were performed once. For gel source data, see Supplementary Fig. 1.



ribosome concentration

$$= \frac{\text{ribosome number}}{\text{cell volume * cell division factor}}$$

ribosome number = $A_0 + A_s - A_d$

A_0 = steaty-state level of r-proteins at a given time

A_s = number of ribosomes synthesized over time

A_d= ribosomes degraded over time

cell division factor = relative cell number increase over time

j

ribosome conc. (0h)

$$\frac{A_0 - 0}{V_0 \ *(1+0/16)} = \boxed{\frac{A_0}{V_0}}$$

ribosome conc. (UT, 12h)

A_s(UT, 12h) = A_0 *(t/16) = 0.75A_0

A_d(UT, 12h) ≈ 0 (relative to Tor1 condition)

cell volume (UT, 12h) = V_0

cell division factor (UT) = (1+ *t*/16)

* one cell division duration = 16h

* *t* = time (h)

$$\frac{(A_0 + A_s - A_d)}{V_0 \ *(1+12/16)} = \frac{A_0 + 0.75A_0}{V_0 * 1.75}$$

$$= \frac{1.75A_0}{V_0 * 1.75} = \boxed{\frac{A_0}{V_0}}$$

*cell division and ribosome synthesis

ribosome conc. (Tor1, 12h)

A_s(Tor1, 12h) = T_r*A_s (UT) = 0.35*A_s (UT)

A_d(Tor1, 12h) = 0.2*A_0

cell volume (Tor1, 12h) = 0.75*V_0

cell division factor (Tor1, 12h) = Y

if one cell division under Tor1 is 24h, Y=1.5

* T_r = translation rate factor

$$\frac{A_0 + 0.35*0.75A_0 - 0.2A_0}{0.75*(V_0)*Y}$$

$$= \frac{1.062*A_0}{0.75*(V_0)*Y} = \boxed{\frac{0.944*A_0}{V_0}}$$

* Y and T_r : variable with cell type, treatment

Predict 5.6% reduction of ribosome concentration after 12h Tor1 treatment, comparable to the 5-8% seen by total proteome analysis.

In autophagy defective cells, A_d(Tor1, 12h) ≈ 0.16*A_0 as autophagy is responsible for 4% of RPS3 degradation out of 20%

Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Systematic decoding of r-protein homeostasis in response to purine deficiency. **a**, Schematic diagram indicating the points of intersection of nucleotides availability with net ribosome production. 6-MP is an inhibitor of purine biosynthesis and blocks production of rRNAs. **b**, 293T cells were treated with or without Tor1, 6-MP, or -Arg medium for the indicated time duration. AHA (250 µM) was added 3 h before collecting the cells. Lysates were either immunoblotted against the indicated antibodies for mTOR signalling inhibition (bottom) or processed for TMR-click reaction for in-gel fluorescence analysis (top). **c**, 293T cells treated with 6-MP for the indicated time points as well as AHA (for the last three hours) were analysed for either mTOR inhibition (middle) or translatome using AHA incorporation assay coupled with TMR-alkyne click (top). Relative translation efficiency is plotted (bottom). Centre data are mean ± s.d. *n* = 1, 3, 3, 3 and 3 biologically independent samples, from left to right. **d**, 293T cells were incubated in the presence or absence of 6-MP for the indicated times, and AHA (250 µM) was added 3 h before collecting each lysate. The translatome was analysed by biotinylation of AHA-labelled proteins followed by TMT-based proteomics. A volcano plot (-Log$_{10}$ p-value versus Log$_2$ 6-MP/UT) showing the translatome and individual r-proteins (in red) at three time points. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 6. *n* = 1 (Neg Ctrl); *n* = 3 (UT, 9h); 2 (6 h, 18 h) biologically independent samples per cell line. **e**, Pre-existing RPL29 in HCT116 RPL29-Halo cells was labelled with TMR (red), washed and then incubated with medium containing green Halo ligand with or without 6-MP (24h). Live cells were imaged. **f**, HCT116 RPS3-Halo cells were subjected to 2-colour labelling as in Fig. 2f. Cells were either left untreated or incubated with 6-MP before analysis of pre-existing and newly synthesized RSP3-Halo using in-gel fluorescence signal-based quantification. Histograms show the relative abundance of pre-existing (red) and newly synthesized (green) RPS3-Halo. Centre data are mean ± s.d. *n* = 2, 3, 3, 3, 3, 3 and 3 biologically independent samples, from left to right for both histograms. **g**, Total proteome analysis of 293T cells (with or without ATG5) was performed according to the scheme (top). Volcano plots (-Log$_{10}$ p-value versus Log$_2$ 6-MP/UT) for all quantified proteins (*n* = 8234 proteins), including individual r-proteins marked with a red dot, are shown at the bottom. *P* values were calculated by two-sided Welch's *t*-test (adjusted for multiple comparisons); for parameters, individual *P* values and *q* values, see Supplementary Table 6. *n* = 3 (UT); 2 (6-MP) biologically independent samples per cell line. **h**, Keima processing assay using the lysates from HEK293 RPS3-Keima WT or *ATG5*$^{-/-}$ cells after 6-MP treatment (24h). Cells treated with arsenite was also blotted as a positive control, as it was previously reported to induce selective ribophagy. **i**, **j**, Analysis of ribosome concentration using biosynthetic, degradative and cell division information is shown as a simple equation in **i**. **j**, Change of ribosome concentration in response to nutrient stress using 0.2 as the degradation rate (derived from AHA-degradomics measurements), translation rates (T$_r$) of 0.35 derived from AHA-translatome analysis, and a cell cycle factor of 1.5 (Y = 1+*t*/24, *t* = 12 h) derived from the proliferation assay. We find that the ribosome concentration upon 12 h of Tor1 treatment [0.944*A$_0$/V$_0$} is comparable to the reduction in ribosomes we measured by total proteome analysis (reduction of ribosomes from ~5–8%). Summary of the systematic quantitative analysis of ribosome inventory during nutrient stress. Related to Fig. 4. Experiments in **b** were performed once, and **e**, **h** were repeated three times independently with similar results. For gel source data, see Supplementary Fig. 1.

nature research

Corresponding author(s): J. Wade Harper

Last updated by author(s): Apr 27, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	1. Orbitrap Eclipse Tribrid Mass Spectrometer (Cat#FSN04-10000) with FAIMS Pro Interface (#FMS02-10001) - Thermo Fisher Scientific 2. Orbitrap Fusion Lumos Tribrid MS (Cat#IQLAAEGAAPFADBMBHQ) with or without FAIMS Pro Interface (#FMS02-10001) - Thermo Fisher Scientific 3. Orbitrap Fusion Tribrid MS (Cat#IQLAAEGAAPFADBMBCX) with FAIMS Pro Interface (#FMS02-10001) - Thermo Fisher Scientific 4. On-line real-time searching was done via Orbiter version 1 (DOI: 10.1021/acs.jproteome.9b00860) a function newly implemented in the latest released Orbitrap Eclipse Tribid Mass Spectrometer factory software (Thermo Fisher Scientific) 5. Odyssey CLx Imager LI-COR bioscience 6. MoxiGo II; ORFLOW Technologies Inc 7. ChemiDocMP (Bio-rad); https://www.bio-rad.com
Data analysis	1. Prism; GraphPad, v7&8 https://www.graphpad.com/scientific-software/prism/ 2. Proteome Discoverer; Thermo Fisher Scientific, v2.3 https://www.Thermo Fisher.com/order/catalog/product/OPTON-30795 3. SEQUEST; Eng et al., (1994) J Am Soc Mass Spectrom. 5 (11): 976–989. 4. Comet (v2018.01 rev. 2); Eng, J.K. et al. (2013), Proteomics 13, 22-24. 5. Perseus; Tyanova et al., Nat Methods. (2016) 13:731-40. http://www.perseus-framework.org 6. FlowJo; V10.5.2 https://www.flowjo.com 7. ImageStudioLite V 5.2.5 https://www.licor.com/bio/products/software/image_studio_lite 8. Image Lab Software V6.0.1 https://www.bio-rad.com 9. FiJi ImageJ V.2.0.0 https://imagej.net/Fiji 10. Microsoft Excel v16 https://www.microsoft.com 11. MetMorph v7.10.1 https://www.moleculardevices.com

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability
All raw mass spectrometry data has been deposited in PRIDE ProteomeXchange (http://www.proteomexchange.org/): Extended Data Set 1 (PXD017852, PXD017853); Extended Data Set 2 (PXD018252); Extended Data Set 3 (PXD017857, PXD018158); Extended Data Set 4 (PXD017855, PXD017856); Extended Data Set 5 (PXD017858, PXD017851); Extended Data Set 6 (PXD017859, PXD017860, PXD017861). Source data for all proteomics-based plots are provided in Supplementary Tables 1-6. All other reagents are available from the corresponding author upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. Sample size was determined to be adequate based on the magnitude and consistency of measurable differences between groups and low observed variability between samples. For proteomics experiments, we chose n=3 or n=4 given the limitation of the available TMT channels. For Flow-cytometry experiments, we analyzed >3000 cells with triplicate experiments which showed consistent results through-out the replicates. In-gel fluorescent analyses were also performed at least three times.
Data exclusions	No data were excluded from the analyses.
Replication	All experiments were replicated and all attempts at replication were successful and consistent. For proteomics experiments replicates clustered together in PCA, and we observed low coefficient of variation among replicates.
Randomization	Proteomics samples for comparison with TMT and TMTpro reagents, were randomly allocated in the TMT/TMTpro group and replicates were in adjacent channels. For other experiments, no randomization was done.
Blinding	Blinding was not relevant in this study, because all the data were analyzed using unbiased methods. Furthermore, the whole image fields were shown in the figures without subjective cropping.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	1. RPS3 Cell Signaling Technology; Cat. No. 9538; RRID:AB_10622028; Dilution 1:2000 2. RPL28 Abcam; Cat. No. ab138125; RRID:AB_10622028; Dilution 1:2000 3. RPS15a Bethyl Lab; Cat. No. A304-990A-T; RRID: AB 2782308; Dilution 1:1000

4. ATG7 Cell Signaling Technology; Cat. No. 8558S; RRID:AB_10831194; Dilution 1:1000
5. Keima MBL international; Cat. No. M182-3; RRID:AB_10794910; Dilution 1:2000
6. LC3B MBL international; Cat. No. M186-3; RRID:AB_10897859; Dilution 1:1000
7. ATG5 Cell Signaling Technology; Cat. No. 12994; RRID:AB_2630393; Dilution 1:1000
8. p70 S6K phospho-T389 Cell Signaling Technology; Cat. No. 9234S; RRID:AB_2269803; Dilution 1:1000
9. phospho-S6 ribosomal protein Ser253, 236 Cell Signaling Technology; Cat. No. 4858; Dilution 1:3000
10. TEX264 Sigma; Cat. No. HPA017739; RRID:AB_1857910; Dilution 1:1000
11. RPL23 Bethyl Lab; Cat. No. A305-010A-T;RRID:AB_2782326; Dilution 1:1000
12. RPL7 Bethyl Lab; Cat. No. A300-741A-T; RRID:AB 2779432; Dilution 1:1000
13. RPL29 Proteintech Group; Cat. No. 15799-1-AP; Dilution 1:2000
14. Tubulin Abcam; Cat. No. ab7291; RRID:AB_2241126; Dilution 1:3000
15. SQSTM1 Novus Biologicals; Cat. No. H00008878-M01; RRID:AB_548364; Dilution 1:2000
16. Anti-puromycin antibody EMD millipore; Cat. No. MABE343; RRID:AB_2566826; Dilution 1:3000
17. NUFIP1 Proteintech Group; Cat. No. 12515-1-AP; RRID:AB_2298759; Dilution 1:1000
18. RPS6 Cell Signaling Technology; Cat. No. 2217; RRID:AB_331355; Dilution 1:3000
19. 4EBP1 Cell Signal Technology; Cat. No. 9644; RRID:AB_2097841; Dilution 1:3000
20. Lamin A/C Cell Signal Technology; Cat. No. 4777; RRID:AB_10545756; Dilution 1:1000
21. TFEB Cell Signal Technology; Cat. No. 4042; Dilution 1:1000
22. YBX1 Proteintech Group; Cat. No. 20339-1-AP; RRID:AB_10665424; Dilution 1:1000
23. Sestrin 2 (SESN2) Proteintech Group; Cat. No. 10795-1-AP; RRID:AB_2185480; Dilution 1:1000
24. SDAD1 Bethyl Lab; Cat. No. A304-692A-T; RRID:AB_2620887; Dilution 1:1000
25. NVL Proteintech Group; Cat. No. 16970-1-AP; RRID:AB_2157811; Dilution 1:1000
26. c-Maf RnD systems; Cat. No. MAB8227-SP; Dilution 1:1000
27. IRDye® 800CW Streptavidin LI-COR; Cat. No. 926-32230; Dilution 1:15000
28. IRDye 800CW Goat anti-Rabbit IgG H+L LI-COR; Cat. No. 925-32211; RRID:AB_2651127; Dilution 1:15000
29. IRDye 680 RD Goat anti-Mouse IgG H+L LI-COR; Cat. No. 926-68070; RRID:AB_10956588; Dilution 1:15000

Validation	Specificity (human) for NUFIP1, ATG5, ATG7, RB1CC1 and TEX264 antibodies were determined by testing CRISPR deleted cells as controls. For RPS3 and RPL29, specificity for immunoblotting (human) was determined by tagging of the respective endogenous gene. The overall immunoblotting results agreed with next-generation sequencing results of NUFIP1, ATG5, TEX264, RPS3, RPL29 engineered cells and quantitative proteomics results of NUFIP1 and TEX264 deleted cells. For other antibodies, we employed the manufacturer's validation provided in individual datasheets.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Human: HEK293 ATCC CRL-1573; RRID:CVCL_0045 Human: HCT116 ATCC CCL-247; RRID:CVCL_0291 Human: HEK293T ATCC CRL-3216; RRID:CVCL_0063
Authentication	Karyotyping (GTG-banded karyotype) of HCT116, 293, and 293T cells (from ATCC) was performed by Brigham and Women's Hospital Cytogenomics Core Laboratory.
Mycoplasma contamination	All cell lines were found to be free of mycoplasma using Mycoplasma Plus PCR assay kit (Agilent).
Commonly misidentified lines (See ICLAC register)	none

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	No tissue processing were used. only cultured HCT116 cells gene edited with Halo tags were used for flow cytometry. Cells were labeled with either TMR (red) ligand or R110 (green) ligand via the Halo enzyme.
Instrument	LSR-II Analyser, BD Biosciences
Software	FlowJoTM; V10.5.2 https://www.flowjo.com
Cell population abundance	Typically >3,000 cells were analyzed, as indicated.

Gating strategy	1. singlet cells were gated by SSC1 hight/FSC1 hight (G1) followed by SSC1 hight/SSC1-width (G2). 2. R110 and TMR signal was measured by 488ex/513(26)em and 561ex/579(16)em and data exported to prism for ratio-metric calculation

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Molecular architecture of the human 17S U2 snRNP

https://doi.org/10.1038/s41586-020-2344-3

Received: 21 November 2019

Accepted: 19 March 2020

Zhenwei Zhang[1,6], Cindy L. Will[2,6], Karl Bertram[1], Olexandr Dybkov[2], Klaus Hartmuth[2], Dmitry E. Agafonov[2], Romina Hofele[3,5], Henning Urlaub[3,4], Berthold Kastner[2], Reinhard Lührmann[2]✉ & Holger Stark[1]✉

The U2 small nuclear ribonucleoprotein (snRNP) has an essential role in the selection of the precursor mRNA branch-site adenosine, the nucleophile for the first step of splicing[1]. Stable addition of U2 during early spliceosome formation requires the DEAD-box ATPase PRP5[2–7]. Yeast U2 small nuclear RNA (snRNA) nucleotides that form base pairs with the branch site are initially sequestered in a branchpoint-interacting stem–loop (BSL)[8], but whether the human U2 snRNA folds in a similar manner is unknown. The U2 SF3B1 protein, a common mutational target in haematopoietic cancers[9], contains a HEAT domain ($\text{SF3B1}^{\text{HEAT}}$) with an open conformation in isolated SF3b[10], but a closed conformation in spliceosomes[11], which is required for stable interaction between U2 and the branch site. Here we report a 3D cryo-electron microscopy structure of the human 17S U2 snRNP at a core resolution of 4.1 Å and combine it with protein crosslinking data to determine the molecular architecture of this snRNP. Our structure reveals that $\text{SF3B1}^{\text{HEAT}}$ interacts with PRP5 and TAT-SF1, and maintains its open conformation in U2 snRNP, and that U2 snRNA forms a BSL that is sandwiched between PRP5, TAT-SF1 and $\text{SF3B1}^{\text{HEAT}}$. Thus, substantial remodelling of the BSL and displacement of BSL-interacting proteins must occur to allow formation of the U2–branch-site helix. Our studies provide a structural explanation of why TAT-SF1 must be displaced before the stable addition of U2 to the spliceosome, and identify RNP rearrangements facilitated by PRP5 that are required for stable interaction between U2 and the branch site.

To our knowledge, at present no high-resolution, cryo-electron microscopy (cryo-EM) structure of the human 17S U2 snRNP, the major subunit of the spliceosomal E and A complexes (Extended Data Fig. 1), is available. We thus determined its 3D structure by cryo-EM (Extended Data Table 1, Extended Data Fig. 2). Consistent with previous low-resolution electron microscopy structures of isolated U2 snRNPs[12], and its overall structure in human spliceosomes[13–16], human 17S U2 exhibits a bipartite structure, with a 3′ and 5′ domain bridged by several density elements (Fig. 1a). The structure of a major part of the 5′ domain could be determined at an overall resolution of 4.1 Å (Extended Data Fig. 2). However, the 3′ domain and connecting bridges, and parts of the 5′ domain are more dynamic and thus less well-resolved (Fig. 1a, Extended Data Fig. 2). Because of the limited resolution in the more dynamic regions of 17S U2, we used an integrated structural biology approach, fitting known X-ray structures or homology models of structured regions of U2 components into the electron-microscopy density map (Extended Data Table 2), in combination with protein crosslinking coupled with mass spectrometry (Extended Data Fig. 3, Supplementary Table 1) and other biochemical data, to generate a pseudo-atomic model for the well-defined regions of the particle (Fig. 1b).

SF3B1 has an open conformation in 17S U2

The 17S U2 5′ domain consists predominantly of the SF3b complex. SF3B3, PHF5A, SF3B5, and $\text{SF3B1}^{\text{HEAT}}$, which comprise the SF3b core, are located in the most well-defined region of the particle (Fig. 1). The super-helical structure of $\text{SF3B1}^{\text{HEAT}}$, which is composed of 20 tandem HEAT repeats (HRs), exhibits an open conformation in isolated SF3b[10] (Extended Data Fig. 1d). In B and B^{act} spliceosomal complexes, in which U2 stably binds the branch site (BS), $\text{SF3B1}^{\text{HEAT}}$ adopts a closed conformation, encompassing the U2–BS helix and binding the BS adenosine (BS-A) in a pocket formed by PHF5A and the HR15–HR17 of SF3B1 (Extended Data Fig. 1d). In 17S U2 and in the isolated SF3b core crystal, $\text{SF3B1}^{\text{HEAT}}$ has a very similar open conformation, except that HR16 is completely structured in 17S U2 (Extended Data Fig. 4a–c). The structural organization of SF3B5, PHF5A and SF3B3 is also highly similar (Extended Data Fig. 4d). Crosslinking suggests that SF3B6 (also known as P14 or SF3b14a) is also located in a similar region in 17S U2 and isolated SF3b (Extended Data Fig. 4e). Thus, the structure of the SF3b core does not change substantially during 17S U2 assembly and, furthermore, the crystal structure of SF3b represents a functionally relevant conformation. SF3B1 is the target of several precursor mRNA

---

[1]Department of Structural Dynamics, MPI for Biophysical Chemistry, Göttingen, Germany. [2]Cellular Biochemistry, MPI for Biophysical Chemistry, Göttingen, Germany. [3]Bioanalytical Mass Spectrometry, MPI for Biophysical Chemistry, Göttingen, Germany. [4]Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany. [5]Present address: AstraZeneca, Gaithersburg, MD, USA. [6]These authors contributed equally: Zhenwei Zhang, Cindy L. Will. ✉e-mail: reinhard.luehrmann@mpi-bpc.mpg.de; hstark1@gwdg.de

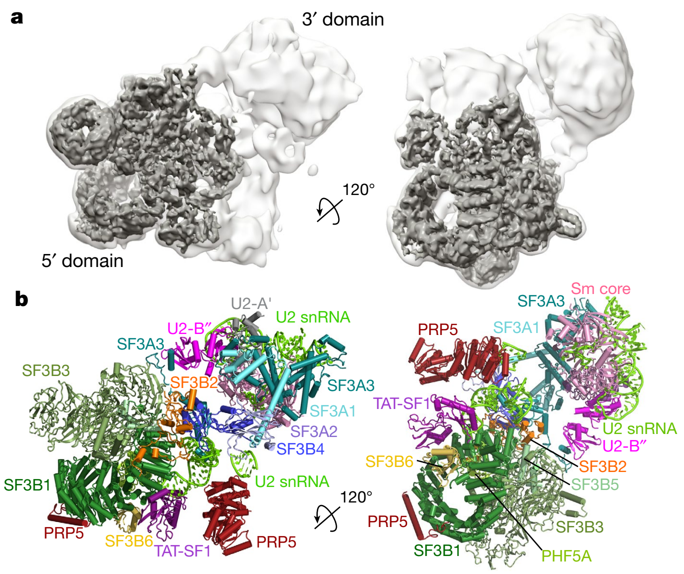


Fig. 1 | 3D cryo-EM model of human 17S U2 snRNP. a, **b**, Different views of the human 17S U2 snRNP cryo-EM density map (**a**) and molecular architecture (**b**). Dark grey denotes better resolved densities.

(pre-mRNA) splicing modulators[9]. The open conformation of SF3B1$^{\mathrm{HEAT}}$ indicates that splicing modulators such as pladienolide B can also interact with it in the 17S U2 snRNP (Extended Data Fig. 4f), and furthermore supports the idea that the SF3B1 conformational change first occurs during stable U2 addition to the spliceosome[17]. The U2 5′ domain is connected by three main bridges to the 3′ domain (Extended Data Fig. 5), which consists of the U2 Sm core RNP, and U2 snRNA stem–loops (SL) III and IV bound by U2-A' and U2-B'' (Fig. 1, Extended Data Fig. 5). Similar to the B and B$^{\mathrm{act}}$ complexes, SF3a proteins, which comprise part of the 3′ domain, have a key role in bridging the U2 Sm core with SF3b in 17S U2[13–16] (Extended Data Fig. 5).

A BSL forms and is bound by protein

In the spliceosome, SF3b contacts SLIIa of U2 snRNA[14,16,18]. On the basis of its length and conformation (Fig. 2, Extended Data Fig. 6a, b), we could unambiguously place U2 SLIIa, as opposed to SLIIb, into one of the two well-resolved helical density elements located close to the C-terminal HEAT repeats of SF3B1 (Extended Data Fig. 6c, d). The high conservation between yeast and humans of U2 nucleotides that form the BSL strongly suggest that the latter also forms in the human U2 snRNP[8]. The location of the helical density that flanks SLIIa indicates that it contains nucleotides 5′ of SLIIa and, as it is a direct continuation of SLIIa, that it accommodates the BSL, as opposed to the mutually exclusive, extended SLI that is separated from SLIIa by 20 nucleotides (Extended Data Fig. 6a, b). Indeed, a modelled BSL fits well in this density (Extended Data Fig. 6e). Moreover, nucleotides 42 to 45 on one BSL strand, which form part of the extended U2–BS helix in the spliceosome, have nearly the same position relative to SLIIa and SF3B1, both in 17S U2 and the spliceosome after U2–BS helix formation (Fig. 2). The BSL is sandwiched between the SF3B1 C-terminal HEAT repeats and other proteins (Fig. 3), and its loop is also sequestered by protein and thus inaccessible for base pairing interactions. Indeed, chaperoning away reactive regions of spliceosomal RNAs and of the pre-mRNA is a common mechanism used by the spliceosome[11,19]. The base of the BSL stem is contacted by a short helix of SF3A3 (designated the separator helix), which ensures an only eight base-pair length of the BSL (Extended Data Fig. 6a). This SF3A3 helix also interacts with SLIIa and its position relative to the latter is maintained in the spliceosome until its catalytic activation[14,16,18] (Extended Data Fig. 6g, h).

PRP5 is near the BSL and encircles SF3B1

Human 17S U2 snRNPs contain TAT-SF1, the function of which in splicing remains unclear, and the DEAD-box ATPase PRP5[6,20]. Prp5 is required for stable U2 addition during formation of the spliceosomal A complex and facilitates a conformational change in U2 snRNP[2–7]. It is also implicated in the displacement of TAT-SF1 (Cus2 in *Saccharomyces cerevisiae*)[5,21], a prerequisite for U2 incorporation into the spliceosome[5,22]. However, the molecular mechanisms by which PRP5 promotes stable U2–BS interaction, and the precise nature of RNP rearrangements that it facilitates, remain poorly understood. In the 17S U2 structure, the PRP5 RecA domains can be fitted into cryo-EM density close to the BSL, U2 snRNA SLIIa and SLIIb, and adjacent to TAT-SF1 (Fig. 3, Extended Data Fig. 7a–c). Consistent with biochemical studies in yeast[23], an extended N-terminal α-helix of PRP5 interacts with SF3B1 HR9–HR12 (Fig. 3a, Extended Data Fig. 7d). Protein crosslinks suggest that PRP5 residues both N- and C-terminal of this α-helix also interact with HR12–HR15 and HR1–HR6, respectively (Fig. 3, Extended Data Fig. 7e). Thus, the N-terminal region of PRP5 seems to encompass SF3B1$^{\mathrm{HEAT}}$, which suggests that it aids in stabilizing its open conformation. The most common (hotspot) cancer-related, point mutations in SF3B1 mainly cluster in or near HR6[10]. Notably, our crosslinking data suggest that the highly conserved PRP5 DPLD motif, which is essential for stable PRP5 interaction with U2 snRNP[24], is located in the proximity of HR6 (Extended Data Fig. 7e), consistent with studies in yeast showing that cancer-related SF3B1 mutations directly destabilize binding of PRP5[23,25].

TAT-SF1 contacts the BSL and SF3B1

RRM1 of TAT-SF1 is located adjacent to SF3B1 HR15–HR16 (Fig. 3, Extended Data Fig. 7f, g), which are thought to act as a hinge[17]. Thus, TAT-SF1 could potentially inhibit their hinge-like function and prevent the closure of SF3B1$^{\mathrm{HEAT}}$ needed to form the BS-A binding pocket.

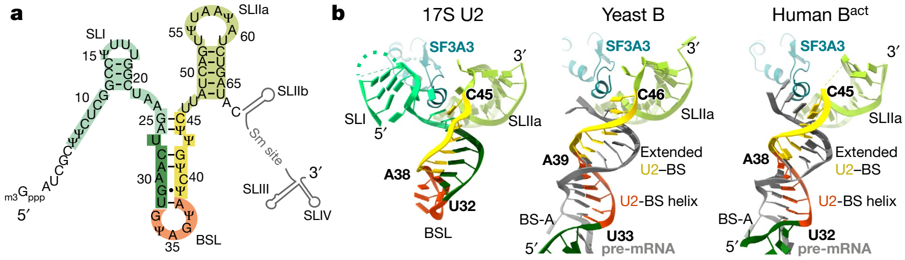


Fig. 2 | A BSL forms in human 17S U2 snRNP. a, 2D structure of human U2 snRNA. SLI nucleotides are in teal; SLIIa nucleotides are in light green. BSL nucleotides that later form the U2–BS helix are in orange, and the extended region of U2–BS helix are in yellow. Remaining BSL nucleotides are in dark green. Ψ denotes pseudouridine. Methylation of U2 bases is not shown (see Extended Data Fig. 6). **b**, 3D structure of U2 SLI, BSL and SLIIa in human 17S U2, and yeast B (PDB code 5NRL) and human B$^{\mathrm{act}}$ (PDB code 6FF4) complexes.

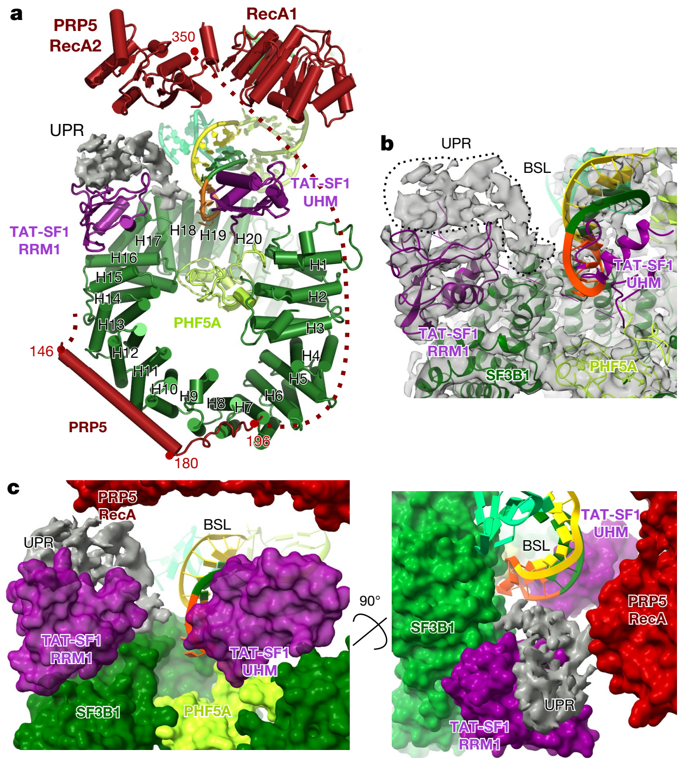


Fig. 3 | PRP5 and TAT-SF1 are located near the BSL and interact with SF3B1$^{\mathrm{HEAT}}$. **a**, Location of the PRP5 RecA1 and RecA2 domains, and interaction of the PRP5 α-helix with SF3B1$^{\mathrm{HEAT}}$, and TAT-SF1 RRM1 and UHM with SF3B1$^{\mathrm{HEAT}}$ and the BSL. Dotted line denotes the potential path of unstructured PRP5 regions, based on crosslinking data. **b**, Fit of TAT-SF1$^{\mathrm{RRM1}}$ into cryo-EM density adjacent to SF3B1 HR15–HR16. A density element that could not be unambiguously defined, but is probably part of TAT-SF1, contacts the BSL loop. **c**, The BSL is sequestered by PRP5, TAT-SF1, SF3B1 and the unassigned protein region (UPR). Grey denotes cryo-EM density of the unassigned protein region. Coloured surfaces are derived from fitted protein models.

TAT-SF1$^{\mathrm{RRM1}}$ is closely associated with a density element that interacts directly with the BSL loop and is contacted by PRP5$^{\mathrm{RecA2}}$ (Fig. 3). This density element is probably part of TAT-SF1, as electron-microscopy density can be traced in a continuous manner from it to density comprising TAT-SF1$^{\mathrm{RRM}}$ (Fig. 3). However, we cannot unambiguously identify its sequence. Regardless of its nature, this protein module helps to stabilize the position of the BSL loop close to SF3B1$^{\mathrm{HEAT}}$. TAT-SF1 contains a U2AF homology motif (UHM) (Extended Data Fig. 7a) that has affinity for the five U2AF ligand motifs (ULMs) present in the SF3B1 N-terminal domain, but preferentially binds to ULM5, located closest to the HEAT domain[26]. Consistent with protein–protein crosslinks, TAT-SF1$^{\mathrm{UHM}}$ can be fitted into a less-well resolved density element near the BSL and PRP5 RecA domains (Fig. 3, Extended Data Fig. 7c, h). Together, the location of PRP5 and TAT-SF1 suggests that they aid in stabilizing both the BSL and SF3B1 open conformation, and thus must be displaced to allow a stable U2–BS interaction.

Model of PRP5 remodelling of U2 snRNP

A comparison of our 17S U2 snRNP structure and the structure of U2 in the spliceosome indicates that a major RNP rearrangement must occur to free those U2 nucleotides that subsequently form the U2–BS helix and allow stable U2 addition during formation of the A complex (Extended Data Fig. 8a). U2 snRNP first associates with the spliceosomal E complex via protein–protein contacts, including those between PRP5 and the U1 snRNP[7], and potentially between U2 proteins and SF1 (also known as MBBP) bound to the BS, that would bring U2 into the vicinity of the BS[27] (Fig. 4). If accessible, BSL loop nucleotides could in principle form base pairs with the BS without disruption of the BSL[8]. However, in human 17S U2, the BSL is sandwiched between the PRP5 RecA domains and TAT-SF1$^{\mathrm{UHM}}$ on one side, and HR16–HR19 of SF3B1 and TAT-SF1$^{\mathrm{RRM1}}$ on the other, with the unassigned protein region directly contacting the BSL loop. PRP5 is a DEAD-box ATPase and thus is expected to bind double-stranded RNA[3,28]. However, the nature of the RNA bound by PRP5 is not clear from our U2 structure and PRP5 may first bind RNA after 17S U2 interacts with the E complex. Several DEAD-box proteins disrupt protein–RNA interactions, in some cases apparently even without unwinding dsRNA[28]. Our data are consistent with a model in which, after U2 incorporation into the E complex, ATP binding would lead to engagement of PRP5 with the BSL or adjacent RNA nucleotides, with hydrolysis leading to displacement or release of BSL-interacting proteins, including TAT-SF1 and the RecA domains of PRP5 itself, without disrupting the BSL (Fig. 4a, b). This would allow initial U2–BS base pairing via the BSL loop, which would also require displacement of SF1 from the pre-mRNA BS nucleotides to free them for base pairing interactions[29]. In yeast, Prp5 ATPase activity is not required if CUS2 (the yeast TAT-SF1 homologue) is no longer present[5]. We thus propose that the BSL, which is no longer stabilized by protein, will subsequently become flexible, and thus that its unwinding and the concomitant formation of the extended U2–BS helix, may not require exogenous energy (Fig. 4b–d). However, at present we cannot exclude that PRP5 destabilizes the BSL in an ATP-dependent manner, and that this leads to the concomitant displacement of TAT-SF1 and other BSL-interacting

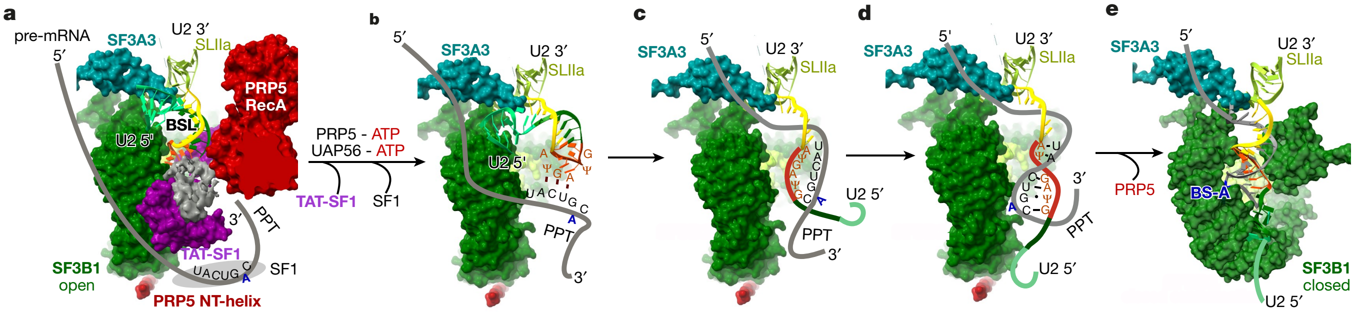


Fig. 4 | Model of PRP5-mediated remodelling events leading to stable U2–BS interaction. **a**–**e**, Model of U2 snRNP rearrangements during the transition from the spliceosomal E to A complex. U1 snRNP and U2AF are omitted for simplicity. **a**, **b**, ATP hydrolysis by PRP5 leads to the removal or displacement of TAT-SF1, the PRP5 RecA domains, and the unassigned protein region (grey) from the BSL, and BSL contact with SF3B1$^{\mathrm{HEAT}}$ is disrupted. This frees the BSL loop for initial base pairing with the BS and leads to destabilization of its stem. **b**, **c**, Movement of the U2 5′ end downward aids BSL unwinding. **c**, **d**, Rotation of the SLI-containing U2 5′ end generates helical turns required to form the extended U2–BS helix. **d**, **e**, A conformational change in SF3B1 clamps the newly formed U2–BS helix within its HEAT domain and leads to PRP5 release. NT-helix, N-terminal helix; PPT, polypyrimidine tract.

proteins. For topological reasons, formation of the extended U2–BS helix would require that the 5′ strand of the BSL no longer contacts the SF3B1 C-terminal HEAT repeats. This would be a prerequisite to free the SLI-containing 5′ end of U2 snRNA for its proposed downward movement (Fig. 4b–d, Extended Data Fig. 8b, c). Rotational movement of the 5′ end of the U2 snRNA would then facilitate formation of the extended U2–BS helix (Fig. 4b–d, Extended Data Fig. 8c). The latter would then move back towards the C-terminal SF3B1 HEAT repeats such that U2 nucleotides 42–45 occupy the same position as they did before U2–BS formation, and the base of the bulged BS-A would be positioned at the hinge region between HR15 and HR17 and close to PHF5A. We propose that docking of the BS-A to the hinge region may trigger several coordinated movements of the HEAT domain leading to its closed conformation and clamping of its terminal HRs onto the extended part of the U2–BS helix (Fig. 4e).

In yeast, Prp5 also functions in an ATP-independent manner during formation of the A complex and remains bound to the latter if the BS is mutated[5,30]. In 17S U2, the N-terminal region of PRP5 makes numerous contacts with SF3B1$^{\mathrm{HEAT}}$ and therefore potentially stabilizes its open conformation. Thus, there may be competition between the maintenance of an open conformation, facilitated by the PRP5 N-terminal region, and the switch to a closed conformation with the BS-A tightly clamped. In this model, only if an appropriate U2–BS helix is formed and the BS-A is properly docked, will the PRP5-SF3B1$^{\mathrm{HEAT}}$ interaction be dissolved, leading to PRP5 release. In this way, PRP5 could potentially proofread stable accommodation of the BS-A during A complex formation. Given the weak conservation of the mammalian BS, further studies are needed to better understand how PRP5 can potentially distinguish between ideal U2–BS duplexes and those containing several mismatches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2344-3.

1. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol*. **3**, a003707 (2011).
2. Ruby, S. W., Chang, T. H. & Abelson, J. Four yeast spliceosomal proteins (PRP5, PRP9, PRP11, and PRP21) interact to promote U2 snRNP binding to pre-mRNA. *Genes Dev*. **7**, 1909–1925 (1993).
3. O'Day, C. L., Dalbadie-McFarland, G. & Abelson, J. The *Saccharomyces cerevisiae* Prp5 protein has RNA-dependent ATPase activity with specificity for U2 small nuclear RNA. *J. Biol. Chem*. **271**, 33261–33267 (1996).
4. Abu Dayyeh, B. K., Quan, T. K., Castro, M. & Ruby, S. W. Probing interactions between the U2 small nuclear ribonucleoprotein and the DEAD-box protein, Prp5. *J. Biol. Chem*. **277**, 20221–20233 (2002).
5. Perriman, R., Barta, I., Voeltz, G. K., Abelson, J. & Ares, M., Jr. ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA. *Proc. Natl Acad. Sci. USA* **100**, 13857–13862 (2003).
6. Will, C. L. et al. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J*. **21**, 4978–4988 (2002).
7. Xu, Y. Z. et al. Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO J*. **23**, 376–385 (2004).
8. Perriman, R. & Ares, M., Jr. Invariant U2 snRNA nucleotides form a stem loop to recognize the intron early in splicing. *Mol. Cell* **38**, 416–427 (2010).
9. Bonnal, S., Vigevani, L. & Valcárcel, J. The spliceosome as a target of novel antitumour drugs. *Nat. Rev. Drug Discov*. **11**, 847–859 (2012).
10. Cretu, C. et al. Molecular architecture of SF3b and structural consequences of its cancer-related mutations. *Mol. Cell* **64**, 307–319 (2016).
11. Kastner, B., Will, C. L., Stark, H. & Lührmann, R. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harb. Perspect. Biol*. **11**, a032417 (2019).
12. Krämer, A., Grüter, P., Gröning, K. & Kastner, B. Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J. Cell Biol*. **145**, 1355–1368 (1999).
13. Bertram, K. et al. Cryo-EM structure of a pre-catalytic human spliceosome primed for activation. *Cell* **170**, 701–713.e11 (2017).
14. Haselbach, D. et al. Structure and conformational dynamics of the human spliceosomal B$^{\mathrm{act}}$ complex. *Cell* **172**, 454–464.e11 (2018).
15. Zhan, X., Yan, C., Zhang, X., Lei, J. & Shi, Y. Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Res*. **28**, 1129–1140 (2018).
16. Zhang, X. et al. Structure of the human activated spliceosome in three conformational states. *Cell Res*. **28**, 307–322 (2018).
17. Cretu, C. et al. Structural basis of splicing modulation by antitumor macrolide compounds. *Mol. Cell* **70**, 265–273.e8 (2018).
18. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017).
19. Papasaikas, P. & Valcárcel, J. The spliceosome: The ultimate RNA chaperone and sculptor. *Trends Biochem. Sci*. **41**, 33–45 (2016).
20. Agafonov, D. E. et al. Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Mol. Cell. Biol*. **31**, 2667–2682 (2011).
21. Talkish, J. et al. Cus2 enforces the first ATP-dependent step of splicing by binding to yeast SF3b1 through a UHM-ULM interaction. *RNA* **25**, 1020–1037 (2019).
22. Xu, Y. Z. & Query, C. C. Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Mol. Cell* **28**, 838–849 (2007).
23. Tang, Q. et al. SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes Dev*. **30**, 2710–2723 (2016).
24. Shao, W., Kim, H. S., Cao, Y., Xu, Y. Z. & Query, C. C. A U1-U2 snRNP interaction network during intron definition. *Mol. Cell. Biol*. **32**, 470–478 (2012).
25. Carrocci, T. J., Zoerner, D. M., Paulson, J. C. & Hoskins, A. A. SF3b1 mutations associated with myelodysplastic syndromes alter the fidelity of branchsite selection in yeast. *Nucleic Acids Res*. **45**, 4837–4852 (2017).
26. Loerch, S. et al. The pre-mRNA splicing and transcription factor Tat-SF1 is a functional partner of the spliceosome SF3b1 subunit via a U2AF homology motif interface. *J. Biol. Chem*. **294**, 2892–2902 (2019).
27. Crisci, A. et al. Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. *Nucleic Acids Res*. **43**, 10456–10473 (2015).
28. Pan, C. & Russell, R. Roles of DEAD-box proteins in RNA and RNP Folding. *RNA Biol*. **7**, 667–676 (2010).
29. Liu, Z. et al. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001).
30. Liang, W. W. & Cheng, S. C. A novel mechanism for Prp5 function in prespliceosome formation and proofreading the branch site sequence. *Genes Dev*. **29**, 81–93 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Affinity purification of 17S U2 snRNPs

HeLa S3 cells were obtained from GBF, Brunswick (currently Helmholtz Center for Infection Research) and tested negative for mycoplasma. To isolate 17S U2 snRNPs, HeLa nuclear extract was prepared as previously described[31]. Extract was loaded onto a 17–30% (w/v) sucrose gradient containing GK5M buffer (20 mM HEPES-KOH, pH 7.9, 150 mM KCl, 5 mM $MgCl_2$, 0.2 mM EDTA, pH 8.0) and centrifuged in a Beckman Ti-15 zonal rotor for 40 h at 32,000 rpm. The gradient was subsequently fractionated, and those fractions comprising the 17S peak were pooled and used for anti-SF3B1 immunoaffinity purification. Under the conditions used, the 17S peak is located approximately in the middle of the gradient, whereas subsequently formed spliceosomal complexes that contain U2 sediment in the 30–45S region, which is near the bottom of the gradient, ensuring that only isolated 17S U2 particles are purified. 17S gradient fractions were diluted 1:1 with GK5M buffer and loaded onto an anti-SF3B1 affinity column. After washing with GK5M buffer containing 1.5% (v/v) sucrose, bound 17S U2 snRNPs were eluted with 0.2 mg ml^{-1} SF3B1 peptide (EQYDPFAEHRPPKIAC) in GK5M containing 1.5% sucrose. Eluates were pooled and separated on a 5–20% (w/v) sucrose gradient containing GK5M buffer and 0–0.15% glutaraldehyde, by centrifuging at 29,000 rpm for 18 h in a Surespin 630 rotor. Gradient fractions were fractionated from the bottom, and glutaraldehyde was quenched by adding 100 mM aspartate (pH 7.0) to each fraction. 17S U2 peak fractions were concentrated and the buffer exchanged with GK5M buffer, by centrifugation with Amicon 50 kDa cut-off units. SDS–PAGE and mass spectrometry indicated that the protein composition of these affinity-purified 17S U2 snRNPs is identical to that of previously described, purified human 17S U2 snRNPs[20]. Owing to the labile nature of the 17S particle and the large excess of U2 snRNP in nuclear extract, it was not possible to efficiently immunodeplete intact 17S U2 snRNPs from HeLa nuclear extract, and assay the function of our purified 17S U2 snRNPs by performing in vitro splicing complementation assays.

BS3 crosslinking of 17S U2 snRNPs and crosslink identification

After elution from the anti-SF3B1 affinity column, purified 17S U2 particles were concentrated to 0.6 μM and subsequently incubated with 60 μM BS3 for 30 min at 20 °C. After quenching by incubating with 10 mM Tris-HCl, pH 8.0, for 15 min at 20 °C, crosslinked 17S U2 particles were loaded onto a 15–30% (w/v) glycerol gradient containing G150 buffer (20 mM HEPES-KOH, pH 7.9, 150 mM KCl, 1.5 mM $MgCl_2$) and centrifuged at 4 °C for 14 h at 35,000 rpm in a TH660 rotor. Gradient fractions were fractionated from the bottom and crosslinked U2 snRNPs migrating in the 17S region of the gradient were pelleted by ultracentrifugation in an S150AT rotor, and analysed as previously described[13]. Peptides were reverse-phase extracted using Sep-Pak Vac tC18 1cc cartridges (Waters) and fractionated by gel filtration on a Superdex Peptide PC3.2/30 column (GE Healthcare). Fifty-microlitre fractions corresponding to an elution volume of 1.2–1.8 ml were analysed in a Thermo Scientific Q Exactive mass spectrometer. Protein–protein crosslinks were identified by pLink 1.23 and 2.3.5 search engines (pfind.ict.ac.cn/software/pLink) and filtered at an FDR of 1% or 5% according to the recommendations of the developer[32,33]. For simplicity, the crosslink score is represented as a negative value of the common logarithm of the original pLink score; that is, score = $-\log_{10}$('pLink Score'). An expected maximum distance between the Cα atoms of the BS3-crosslinked lysine residues is approximately 30 Å. The length of most crosslinks (93–95% at the spectral level and 85–86% at the unique crosslink level) is ≤30 Å in the presented model of the 17S U2 snRNP (see Extended Data Fig. 3). The CXMS data from Supplementary Table 1 can be downloaded and visualized by interactive 2D viewers such as xiNET http://crosslinkviewer.org/ or https://xvis.genzentrum.lmu.de/login.php, or in 3D using UCSF Chimera or PyMOL.

Electron-microscopy sample preparation and image acquisition

The purified 17S U2 sample was absorbed to a thin layer carbon film for 1 min, which was subsequently attached to R3.5/1 QUANTIFOIL grids. Four microliters of ddH_2O was applied to the grid and excess water was blotted away by an FEI Vitrobot loaded with pre-wet filter paper, with the setting of blotting force 8, blotting time 4 s, at 100% humidity and 4 °C, and then vitrified by plunging into liquid ethane. Cryo-EM data were acquired on a FEI Titan Krios electron microscope (Thermo Fisher Scientific), equipped with a Cs corrector, at 300 kV. The images were recorded in integration mode on a Falcon III direct electron detector at 120,700× magnification, which corresponds to a calibrated pixel size of 1.16 Å at the specimen level. Micrographs were recorded using an exposure time of 1 s and a total dose of 72 e^- per $Å^2$. In total, 12,194 were recorded with 20 movie frames and another 28,226 micrographs were recorded with 39 frames.

Image processing

Frames were dose-weighted, aligned and summed with MotionCor2[34]. The defocus values of the micrographs were determined by Gctf[35]. Summed micrographs were manually evaluated in the COW-MicrographQualityChecker (http://www.cow-em.de). Micrographs with isotropic Thong rings and clear particles were processed further. In total, 27,890 out of 40,420 summed micrographs were retained for further processing. Initially, around 3.5 million particles were automatically picked using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang), then extracted with a box size of 360 × 360 pixels, and binned to 180 × 180 pixels (pixel size of 2.32 Å) in RELION 3.0 (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page). Several iterations of reference-free 2D classification were performed in RELION 3.0 and 'bad classes' showing fuzzy or uninterpretable features were removed, yielding around 1.43 million 'good particles'. A subset of 200,000 particles was used to generate an initial 3D map by ab initio reconstruction function in cryoSPARC[36]. The 3′ domain and other peripheral parts of this map were erased in UCSF Chimera v.1.13.1[37]. The remaining 5′ domain was low-pass filtered to 40 Å and used as the 3D reference for 3D classification in RELION 3.0. The approximately 1.43 million good particles from 2D classification were split into five datasets. Each dataset was 3D classified into three classes. All 612,445 particles from good 3D classes (classes that show a complete particle with fine details) were combined and subjected to the Refine3D function in RELION 3.0, with a mask around the 5′ domain, resulting in a 5.7 Å resolution map. Next, using the alignment parameters from the aforementioned masked 3D refinement, the 612,445 particles were focused classified with a mask around the 5′ domain, into 8 classes. The best class (with 152,078 particles) that showed clear separation of HEAT repeats was selected, re-extracted in original pixel size (1.16 Å per pixel) with the box size of 300 × 300 pixels, and subjected to another round of focused 3D classification. The best 3D class (with 120,070 particles) was then selected and refined by Refine3D in RELION 3.0 with a mask around the entire 17S U2 particle or the 5′ domain. The final reconstruction of the entire particle has an average resolution of 7.1 Å and the reconstruction with a mask around 5' domain has an average resolution of 4.1 Å, based on the RELION gold-standard Fourier shell correlation. The data were acquired at two different settings (see above). Because higher averaged doses tend to result in poorer reconstruction, we checked whether removing the particles from the micrographs receiving a higher average dose (leading to 82,420 particles retained) would improve the numerical resolution and the map quality, but this was not the case (data not shown).

Model building and refinement

Templates for the U2 proteins and RNA were obtained, wherever possible from published structures (Extended Data Table 2). The SF3b core crystal structure, together with parts of SF3A3 (amino acids 392–499) and SF3B2 (amino acids 458–530, 565-598) were initially docked as rigid bodies into the 4.1 Å map of the U2 5′ domain. WD40-B of SF3B3 was then manually adjusted to fit the map using Coot v. 0.8.9.2[38]. The central part of SF3b core has resolution ranging from 3.6 to 4.5 Å, which allowed manual adjustment of side chains. The other parts were locally adjusted, without manipulating secondary structures or side chains. The model of the stem of U2 SLI (nucleotides 12–14, 19–21) was built by UCSF Chimera v. 1.13.1[37] as an A-form RNA helix. The model of BSL (nucleotides 25–45) was predicted by RNAvista[39] with base-pairing restrains as proposed in a previous study for *S. cerevisiae* U2 snRNA[8]. U2 SLIIa (nucleotides 48–65) was modelled based on the SLIIa in the yeast $\mathrm{B}^{\mathrm{act}}$ complex, and adjusted to the human U2 snRNA sequence. These RNA models were docked into the map individually as rigid bodies and connections between the SLs were de novo built using Coot. We also built homology models for PRP5 (amino acids 146–195) and TAT-$\mathrm{SF1}^{\mathrm{RRM1}}$ (amino acids 127–220), using SWISS-MODEL suite[40] or SpliProt3D database[41] and guided by our crosslinking (CXMS) data, we then rigidly docked the models into the high resolution map of the 5′ domain. The better resolution (approximately 5 Å) of the TAT-$\mathrm{SF1}^{\mathrm{RRM1}}$ region allowed local adjustment of the homology model. Models of the entire 5′ domain were then combined and subjected to real space refinement in PHENIX[42], with secondary structure restraints. The models of the remaining proteins or RNA (Extended Data Table 2) were rigid body docked into the low-resolution map of the entire 17S U2 particle, based on CXMS data and the overall shape of the EM density.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinate files have been deposited in the Protein Data Bank (PDB) with the following accession codes: U2 5′ domain (6Y50), low resolution region (6Y53) and entire 17S U2 particle (6Y5Q). The cryo-EM maps have been deposited in the Electron Microscopy Data Bank as follows: U2 5′ domain (EMD-10688) and entire U2 particle (EMD-10689).

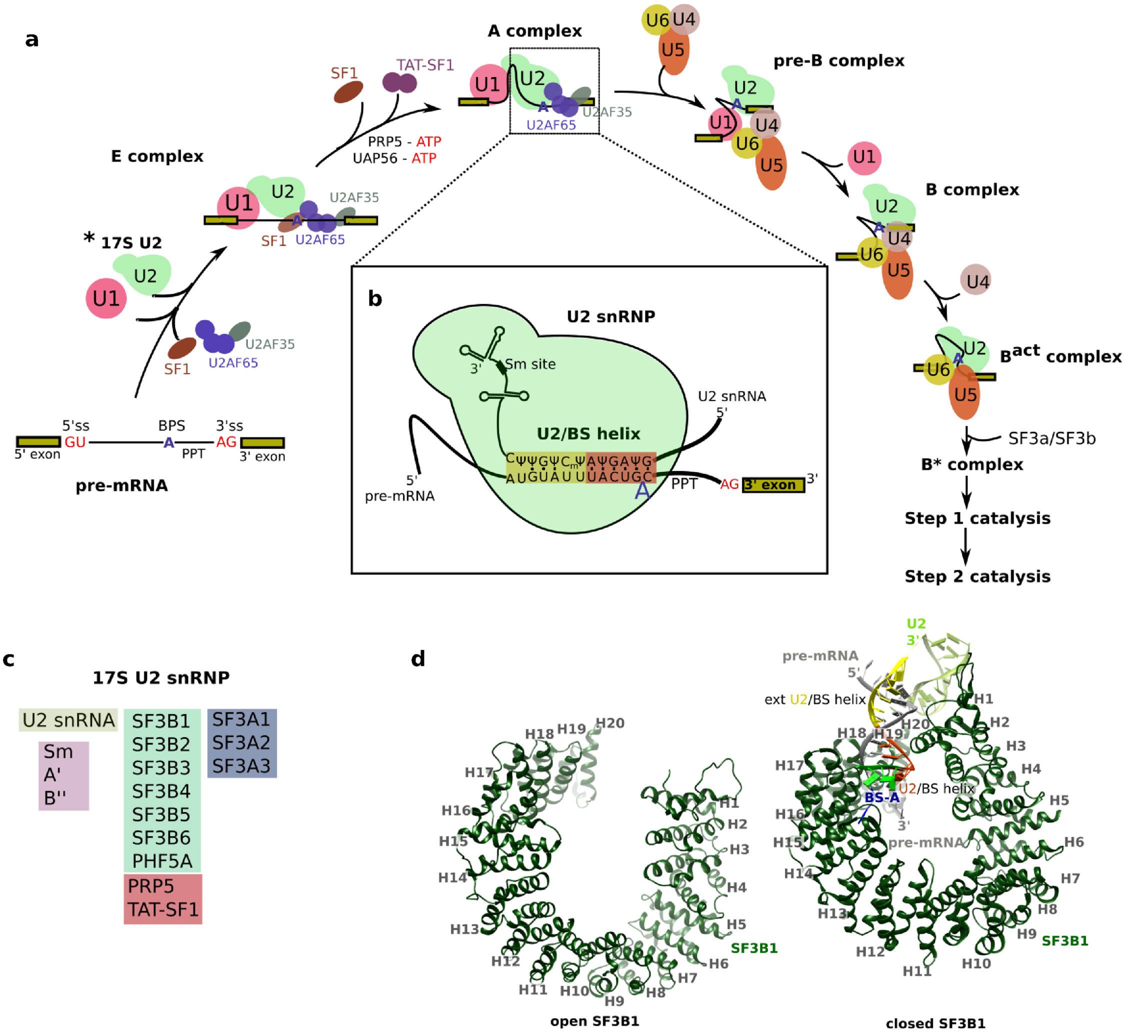
31. Agafonov, D. E. et al. Molecular architecture of the human U4/U6. U5 tri-snRNP. *Science* **351**, 1416–1420 (2016).
32. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
33. Chen, Z. L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun*. **10**, 3404 (2019).
34. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
35. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol*. **193**, 1–12 (2016).
36. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
37. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
39. Antczak, M. et al. RNAvista: a webserver to assess RNA secondary structures with non-canonical base pairs. *Bioinformatics* **35**, 152–155 (2019).
40. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
41. Korneta, I., Magnus, M. & Bujnicki, J. M. Structural bioinformatics of the human spliceosomal proteome. *Nucleic Acids Res*. **40**, 7046–7065 (2012).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Behrens, S. E., Tyc, K., Kastner, B., Reichelt, J. & Lührmann, R. Small nuclear ribonucleoprotein (RNP) U2 contains numerous additional proteins and has a bipartite RNP structure under splicing conditions. *Mol. Cell. Biol*. **13**, 307–319 (1993).
44. MacMillan, A. M. et al. Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev*. **8**, 3008–3020 (1994).
45. Will, C. L. et al. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J*. **20**, 4536–4546 (2001).
46. Spadaccini, R. et al. Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA* **12**, 410–425 (2006).
47. Plaschka, C., Lin, P. C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419–422 (2018).
48. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5′ splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).
49. Darman, R. B. et al. Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. *Cell Rep*. **13**, 1033–1045 (2015).
50. Alsafadi, S. et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat. Commun*. **7**, 10615 (2016).
51. Zhang, J. et al. Disease-causing mutations in SF3B1 alter splicing by disrupting interaction with SUGP1. *Mol. Cell* **76**, 82–95.e7 (2019).
52. Kielkopf, C. L., Lücke, S. & Green, M. R. U2AF homology motifs: protein recognition in the RRM world. *Genes Dev*. **18**, 1513–1526 (2004).

Acknowledgements We thank T. Conrad for the production of HeLa cells in a bioreactor, H. Kohansal for preparing HeLa cell nuclear extract, and G. Heyne, U. Steuerwald, W. Lendeckel, M. Raabe and U. Pleßmann for technical assistance. This work was supported by the Max Planck Society and the Deutsche Forschungsgemeinschaft (SFB 860 grants to H.U., R.L. and H.S.).

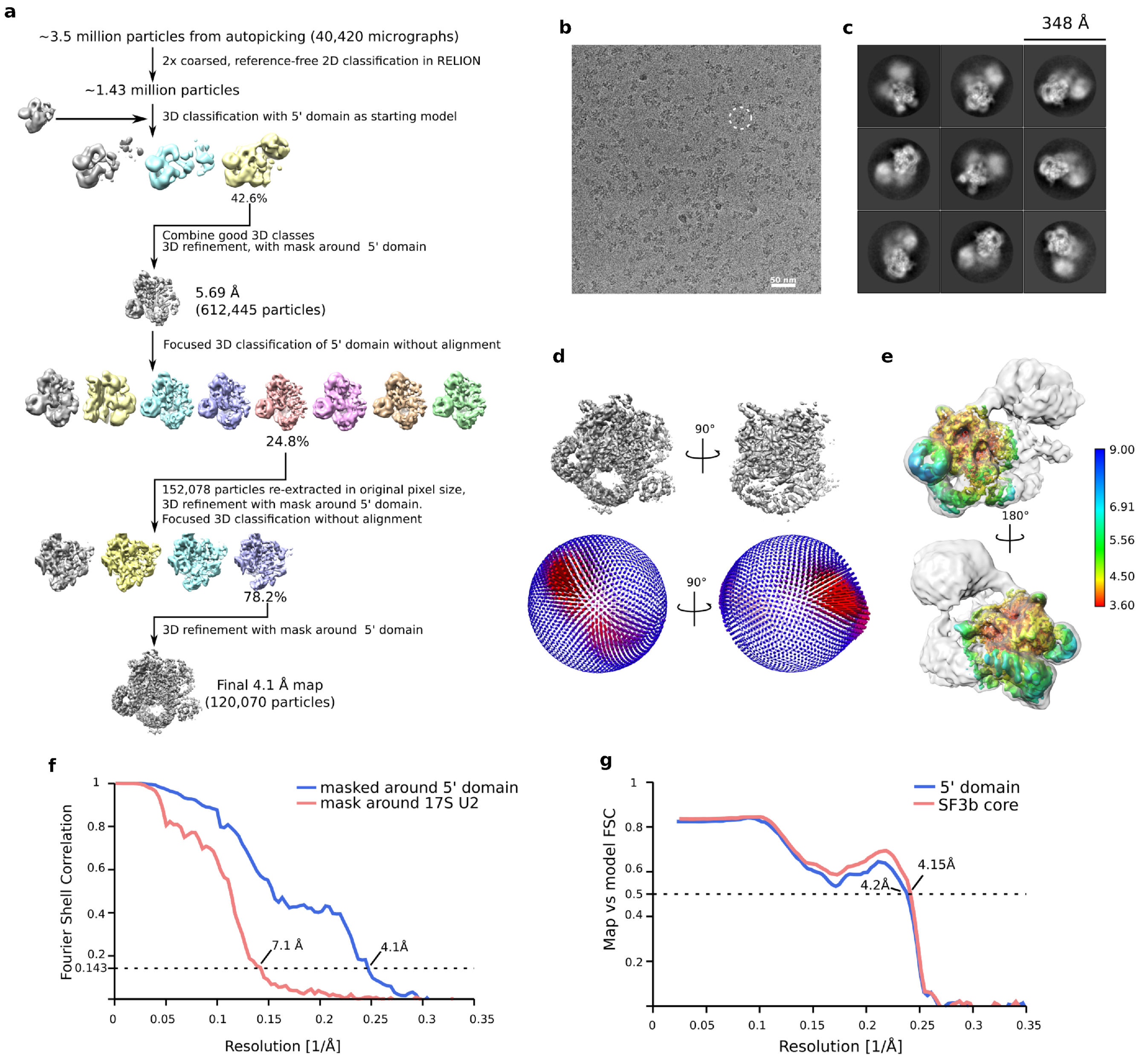
Author contributions C.L.W. and D.E.A. developed the purification strategy and characterized the 17S U2 snRNP. O.D., R.H. and H.U. analysed protein–protein crosslinking data. Z.Z. prepared grids, collected electron microscopy data and determined the structure. Z.Z., B.K., K.H., K.B., R.L. and H.S. analysed the structure. All authors were involved in data interpretation. The manuscript was written by C.L.W. and R.L., with input from all authors. R.L., C.L.W. and H.S. initiated and orchestrated the project.

Competing interests The authors declare no competing interests.

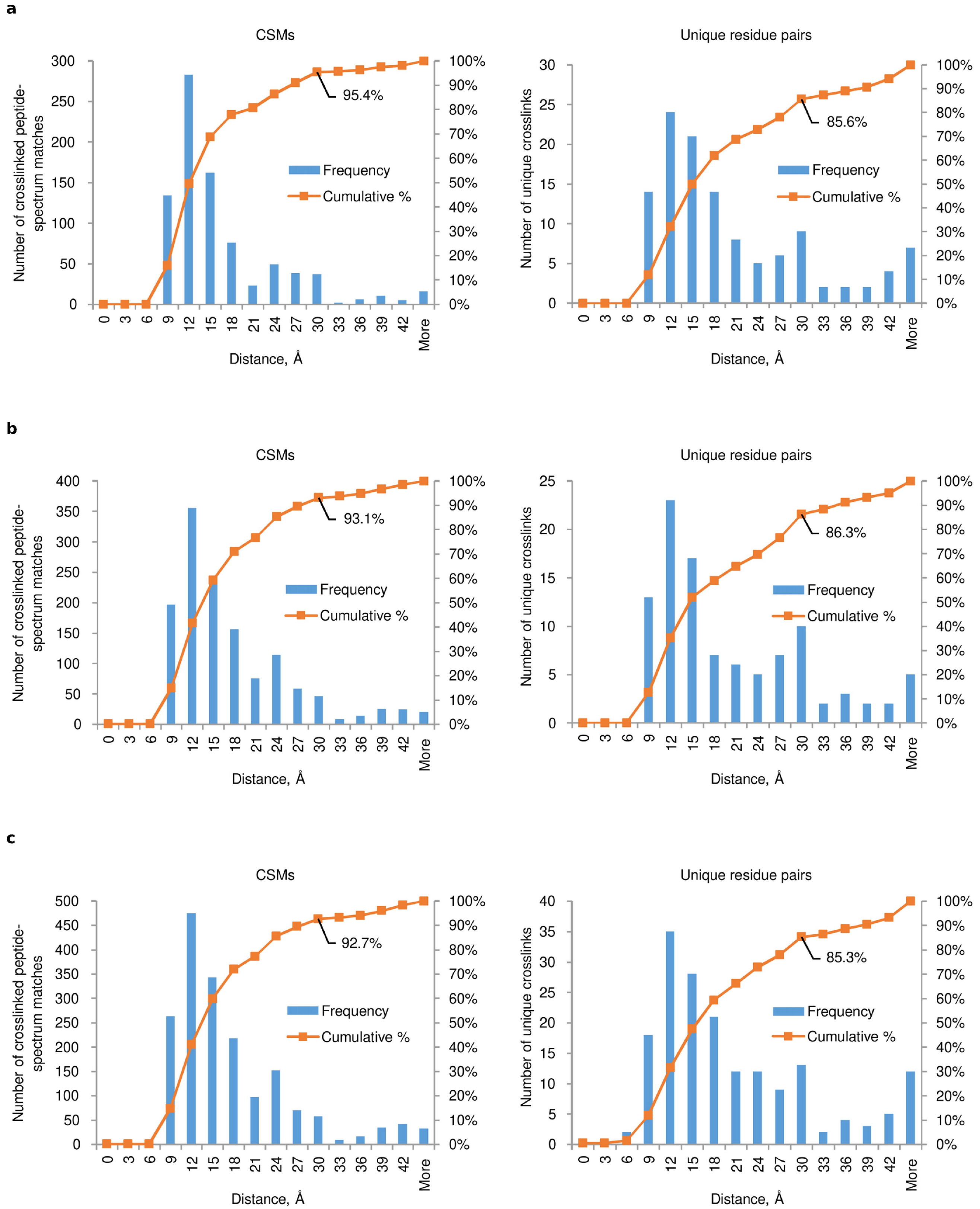
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2344-3.
Correspondence and requests for materials should be addressed to R.L. or H.S.
Peer review information *Nature* thanks Aaron Hoskins, Patrick Schultz, Jonathan Staley and the other, anonymous reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



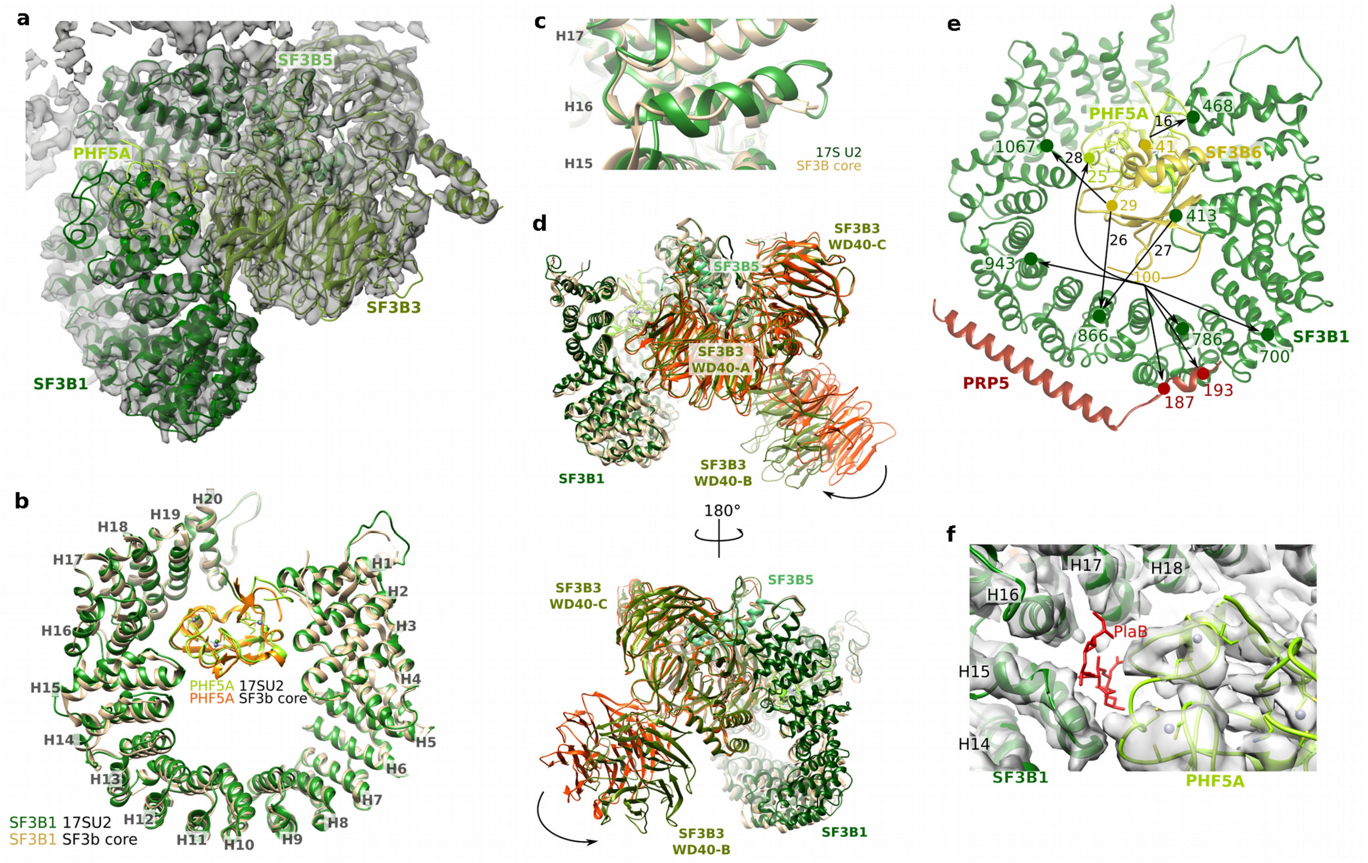
Extended Data Fig. 1 | Spliceosome assembly cycle and the interaction of 17S U2 with the pre-mRNA branch site. a, Early assembly and catalytic activation pathway of the spliceosome. The 17S U2 snRNP, the structure of which was determined here by cryo-EM, is indicated by an asterisk. For simplicity, the stepwise interactions of the U1, U2, U4/U6 and U5 snRNPs, and only selected non-snRNP proteins are shown. Helicases involved in conversion of the E to A complex are indicated. In the E complex, the U2AF65 UHM interacts with the ULM of SF1, and after release of SF1, it subsequently interacts with a ULM in the N-terminal region of SF3B1. This swap of UHM–ULM interactions is probably very important for positioning the BS before the conformational change in the HEAT domain clamps down on the U2–BS helix and stabilizes the U2 snRNP interaction with the intron. The U2AF65/U2AF35 dimer is released (not shown) during conversion of the A to B complex[20]. SF1 is displaced from the BS by UAP56 (either before or after PRP5 action). SF1 pre-bulges the BS-A via accommodation of the latter in its KH domain, facilitating subsequent base-pairing of U2 with the BS[29]. **b**, Base-pairing interactions between U2 snRNA and the BS and upstream intron nucleotides, that lead to bulging of the BS-A. The sequence shown is from intron 10 of the pre-mRNA for the polypyrimidine tract-binding protein (PTB). Red shading denotes the bona fide U2–BS helix, and shows base pairs formed between human U2 snRNA and the conserved BS consensus sequence of PTB intron 10. Yellow shading denotes the extended U2–BS helix, in which the number and nature of base-pairing interactions varies depending on the pre-mRNA intron sequence. **c**, Schematic of the composition of the human 17S U2 snRNP. Only abundant U2 proteins are shown[20]. **d**, Open structure of the SF3B1 HEAT domain in the isolated SF3b complex (left) and its more closed conformation after interaction with the U2–BS helix (right). The SF3B1 HEAT domain (green) forms a super-helical structure, and in the spliceosome sequesters the U2–BS helix. The conformational change in the HEAT domain, which is required to form the BS-A binding pocket, was proposed to occur after formation of the U2–BS helix[17]. Before this study, the conformation of SF3B1 in human 17S U2 snRNP was not known. The pre-mRNA is in grey; U2 snRNA is coloured as in Fig. 2. For simplicity, the PHF5A protein that also forms part of the BS-A binding pocket is not shown. The structures of $\text{SF3B1}^{\text{HEAT}}$ in the isolated SF3b complex (PDB code 5IFE) (left) and human B^{act} complex (PDB code 6FF4) (right) are aligned via HEAT repeat 20.



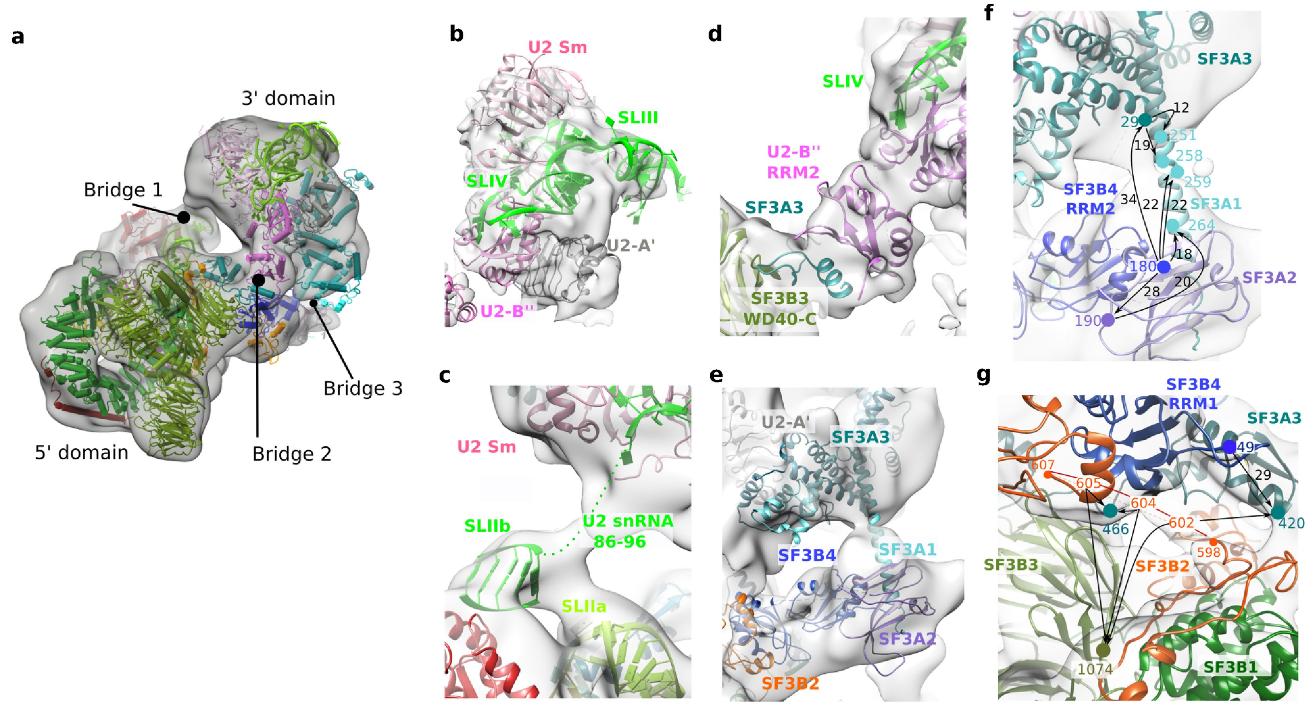
Extended Data Fig. 2 | Cryo-EM and image-processing of the human 17S U2 snRNP. a, Computation sorting scheme. All major image-processing steps are depicted. For a more detailed explanation, see 'Image processing' in the Methods. A considerable amount of conformational heterogeneity is present in all spliceosomal complexes but even more in the bipartite 17S U2 snRNP, which is structurally very labile and readily dissociates during purification[43], making its analysis by electron microscopy challenging. In addition, the bridges connecting the 5′ and 3′ domains of the U2 particle (see also Extended Data Fig. 5) have a very flexible character, leading to flexibility in the 3′ domain and the large variation in local resolution. Thus, a substantially higher number of particles was needed to generate the 17S U2 structure than is usually used for cryo-electron microscopy. **b**, Typical cryo-electron micrograph of the *Homo sapiens* 17S U2 snRNP recorded at 120,700× magnification with a Titan Krios microscope using a Falcon III direct electron detector operating in integration mode at a calibrated pixel size of 1.16 Å. **c**, Representative selection of reference-free 2D class average images depicting 17S U2 particles recorded under cryo conditions. **d**, Euler angle distribution plot of all 17S U2 particles that contributed to the final structure. Red depicts a higher relative number of particles at a certain angle. The generally uniform distribution of the particle projection angles ensures an isotropic 3D electron-microscopy density map. **e**, Local resolution estimation of the 5′ domain of 17S U2 snRNP. The 5′ domain shows a resolution between 3.6 and 9.0 Å. The map of the remaining part excluding the 5′ domain, shown as a translucent overlay, was determined at resolutions between 10 and 30 Å. **f**, Fourier shell correlation (FSC) of two independently refined half datasets, calculated using the 'PostProcessing' routine in RELION, indicates a global resolution of 7.1 Å for the entire 17S U2 snRNP, and 4.1 Å for the masked U2 5′ domain. Multibody refinement around the 3′ domain and the peripheral parts did not produce better resolved maps for those regions. **g**, Map versus model FSC curves for the 5′ domain and the SF3b core. The FSC = 0.5 criterion indicates a resolution of 4.2 Å for the U2 5′ domain, and 4.15 Å for the SF3b core.



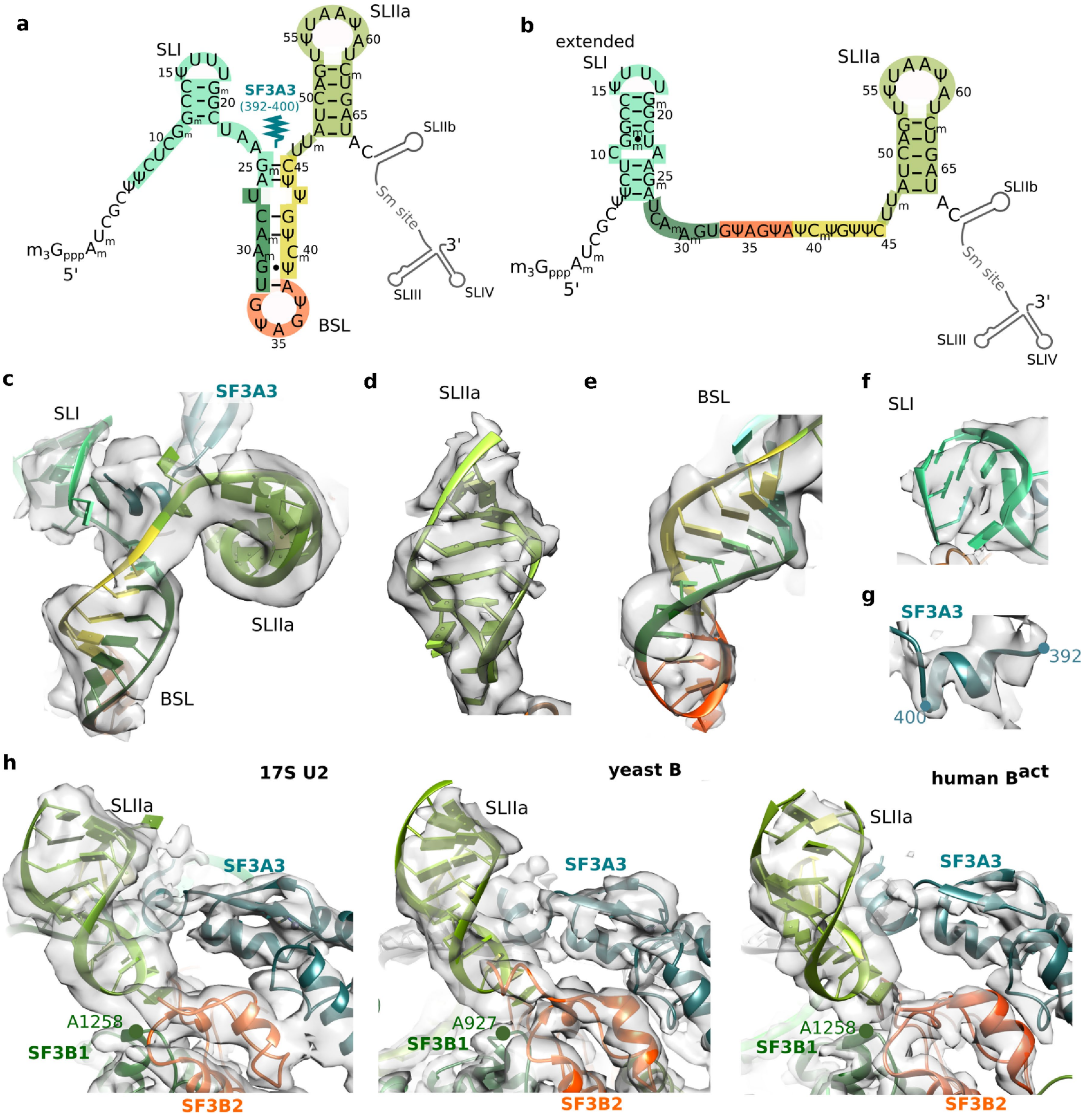
Extended Data Fig. 3 | Euclidian distances for crosslinks observed between modelled residues (Cα) of the 17S U2 snRNP. a–c, Crosslinks from a single 17S U2 crosslinking experiment (with two technical replicates) were identified by pLink2.3.5 and filtered to a false discovery rate (FDR) of 1% (**a**), pLink1.23 at an FDR of 1% (**b**), and pLink1.23 at an FDR of 5% (**c**). Calculations were performed using PyMOL2.3.4 for crosslinks with a score of at least 1. Most crosslinks (93–95% at the spectral level and 85–86% at the unique crosslink level) are consistent—that is, Cα atoms of the crosslinked amino acids are within 30 Å of each other—in the presented model of the 17S U2 snRNP. The percentage of overlength crosslinks (that is, longer than 30 Å) is slightly higher than observed for more rigid complexes, which is consistent with the known structural flexibility/dynamics of the 17S U2 snRNP.



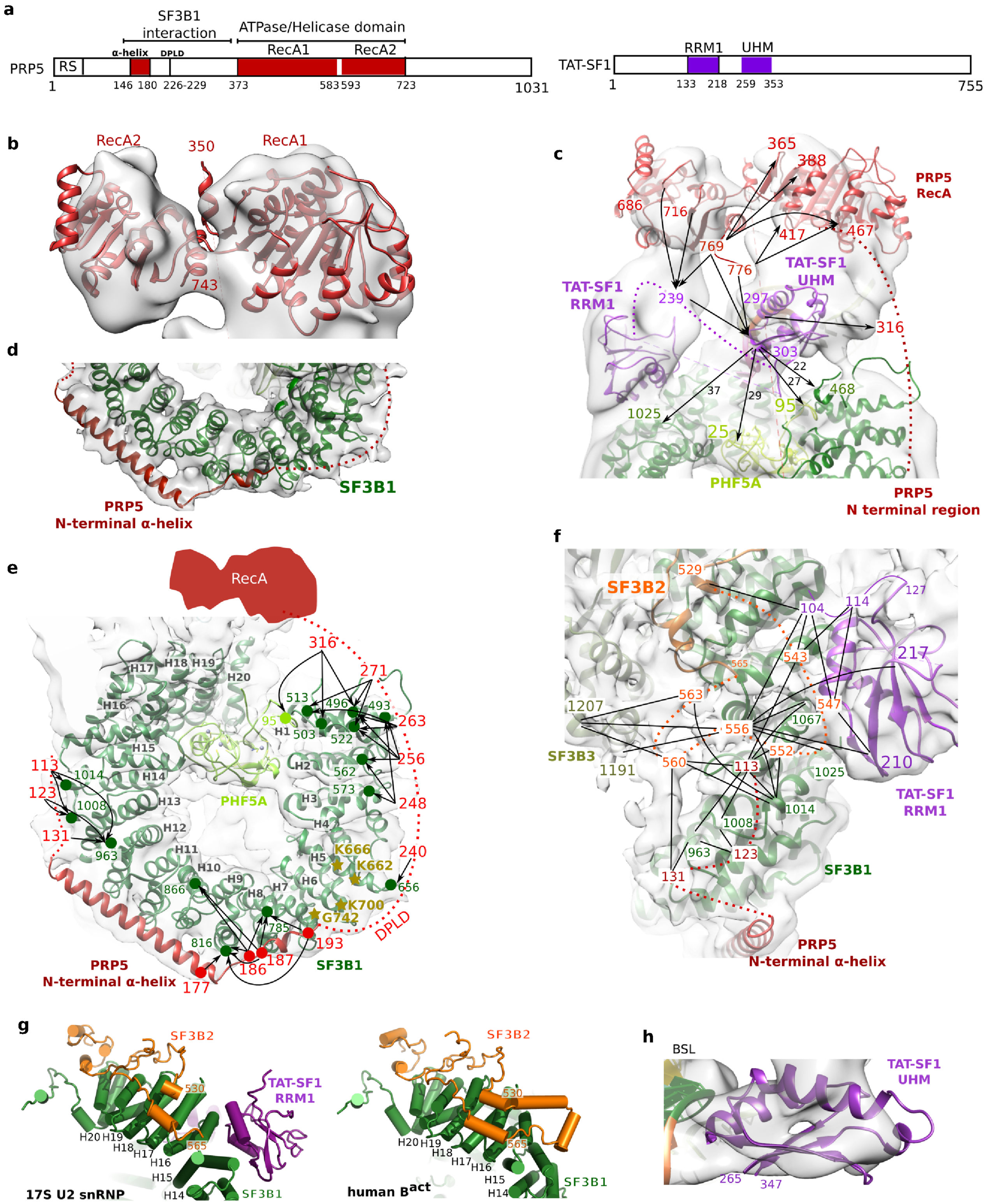
Extended Data Fig. 4 | The SF3B1 HEAT domain has an open conformation in the 17S U2 snRNP. a, Fit of the SF3b core proteins into the 17S U2 electron-microscopy density (grey). **b**, Overlay of the HEAT domain (amino acids 529–1201) of SF3B1 in 17S U2 snRNP (green) and the crystal structure of the isolated SF3b core (gold, PDB 5IFE). In 17S U2 and isolated SF3b, PHF5A contacts both N- and C-terminal regions of the HEAT domain, interacting with HEAT repeats HR2–HR3 near its N terminus, as well as HR15, HR17 and HR18 near its C terminus. PHF5A thus contacts two previously described, dynamic hinge regions (HR3–HR4 and HR15–HR16) of the HEAT domain[10], and thereby helps to stabilize the SF3B1 open conformation. **c**, Close up of SF3B1 HR16 in 17S U2 overlaid with that in isolated SF3b. HR16 is completely structured in the 17S U2 particle, but not in isolated SF3b. **d**, Overlay of the SF3b core domain in 17S U2 snRNP (green) with the crystal structure of isolated SF3b (gold; PDB code 5IFE). For clarity, the SF3B3 protein is coloured red-orange in this panel. Although the WD40-A and WD40-C domains of SF3B3 have essentially the same conformation, and clamp SF3B5 in a similar manner in both the 17S U2 snRNP and the SF3B core complex, the WD40-B domain has a slightly different position and is rotated more towards SF3B1$^{\mathrm{HEAT}}$ in 17S U2. **e**, Multiple crosslinks were detected between SF3B6, which can be crosslinked to the BS-A in spliceosomal A complexes[44,45], and SF3B1 residues on the upper surface of the HEAT domain, as well as the PRP5 α-helix that interacts with HR9–HR12, and PHF5A (see also Supplementary Table 1). Crosslinks between SF3B1 and SF3B6 were also detected in the N-terminal region of SF3B1 located at, or near, amino acids (373–415) that are required for stable SF3B6–SF3B1 interaction[45,46]. Similar protein–protein crosslinks involving SF3B6 were observed with recombinant, intact SF3b complexes[10], which indicates that SF3B6 is located in a similar, but not firmly fixed, position both in 17S U2 and the isolated SF3b complex[10]. Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The distances between the crosslinked residues in our 17S U2 model are indicated by small numbers next to the black arrows. A distance is not included if one of the crosslinked residues is present in an unstructured protein region. **f**, The site where the splicing modulator pladienolide B (PlaB) binds SF3B1 is not occupied in the 17S U2 snRNP. The crystal structure of PlaB bound to the SF3b core complex showed that the binding pocket of PlaB, which is formed by HR15–HR17 and PHF5A, is present only in the open conformation of the HEAT domain, and overlaps with the BS-A-binding pocket[17]. As the PlaB-binding pocket is present in a hinge region of the HEAT domain, it was proposed to inhibit SF3b function by preventing the conformational change in the HEAT domain needed to clamp down on the U2–BS helix[17]. There is no electron-microscopy density observed in the PlaB-binding pocket and thus it and potentially other splicing modulators can also bind SF3B1 in the 17S U2 snRNP.



Extended Data Fig. 5 | The 3′ domain of the 17S U2 snRNP and molecular bridges connecting it to SF3b. a, The U2 3′ domain is connected to the 5′ domain by three main bridges. Fit of the entire 17S U2 molecular model into the electron-microscopy density (low-pass filtered). **b**, Fit of the U2 Sm core, U2 snRNA SLIII, and U2-A' and U2-B″ bound to U2 snRNA SLIV. The overall structure of the U2 3′ domain does not change substantially after U2 incorporation into the spliceosome. The U2 Sm core domain is located at a similar distance from SF3B1$^{\text{HEAT}}$ as observed in the human B complex[13,15]. The resolution of the 17S U2 present in the yeast A complex[47], as well as in the human pre-B complex[15,48], is not sufficient to make meaningful, detailed comparisons of their structure with that of our 17S U2 snRNP. Furthermore, in the former complexes the molecular architecture of U2 is derived entirely from that found in yeast B or human B$^{\text{act}}$ complexes. **c**, Bridge 1 is probably composed of U2 snRNA nucleotides upstream of the Sm-binding site that connect it to SLIIb, which is also part of this bridge, as well as unassigned protein density. **d**, Bridge 2 is formed by RRM2 of U2-B″ and amino acids in the C-terminal half of SF3A3 that bridge U2-B″ and the WD40-C domain of SF3B3. **e**, The N-terminal helical domain of SF3A3 contacts the U2 Sm core. SF3A3 also interacts with SF3B2 and then extends to the U2 snRNA SLIIa and BSL (see Fig. 2). Rigid-body fitting combined with protein–protein crosslinking (see **f** and **g**) allowed us to localize near or within bridge 3, both RRM1 (amino acids 13–91) and RRM2 (amino acids 100–179) of the SF3B4 protein. SF3B4$^{\text{RRM1}}$ interacts with a short region of its binding partner SF3B2 (amino acids 607–693) and SF3B4$^{\text{RRM2}}$ extends towards the β-sandwich of SF3A2 (amino acids 118–209) and a long helical region of SF3A1 (amino acids 235–274) that also comprises part of bridge 3. SF3A1 extends from the N-terminal region of SF3A3 to the SF3A2 β-sandwich. Thus, SF3a proteins have important bridging roles in connecting the 3′ and 5′ domains of the U2 snRNP. Stable integration of SF3a into the human 17S U2 particle during its biogenesis requires the prior binding of SF3b[12] and thus the SF3a–SF3b protein contacts described above potentially have key roles in the assembly of the 17S U2 particle. **f**, **g**, Intermolecular crosslinks supporting the location in our 17S U2 model of SF3B4$^{\text{RRM2}}$ (**f**) and SF3B4$^{\text{RRM1}}$ (**g**). Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The distance between the crosslinked residues in our 17S U2 model is indicated by small numbers next to the black arrows. A distance is not included if one of the crosslinked residues is present in an unstructured protein region.

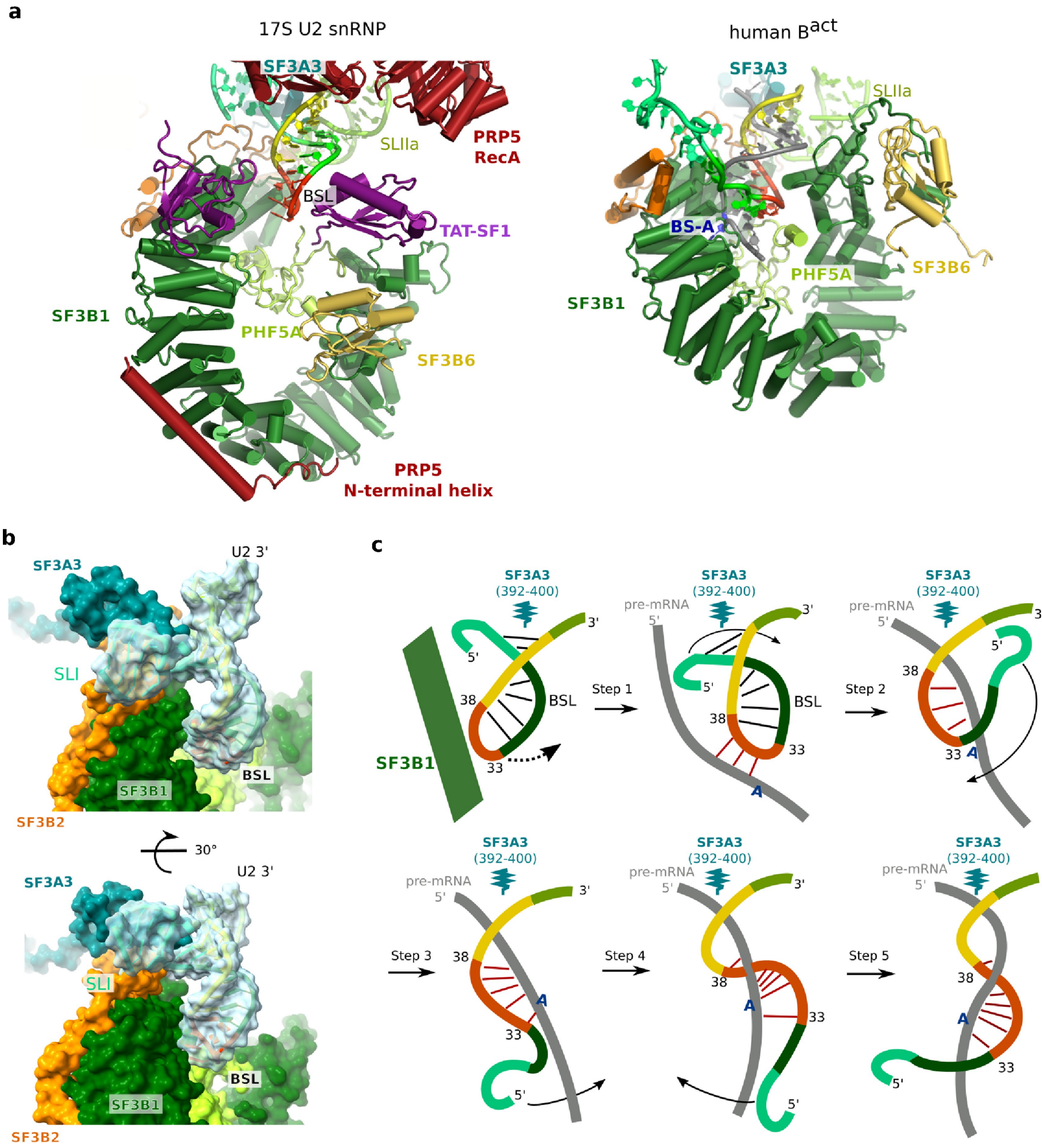


Extended Data Fig. 6 | Fit of the various stem–loops in the 5′ half of U2 snRNA. a, **b**, Alternative conformations of stem–loops potentially formed in the 5′ part of human U2 snRNA. The stem of the BSL could potentially form additional base pairs, but the presence of the SF3A3 separator helix (amino acids Y392 to H400) prevents base-pair formation beyond G25–C45. This guarantees that U46 and U47 are single-stranded and thus could potentially be involved in the proposed movement of the BSL away from SF3B1 during U2–BS helix formation (see also Extended Data Fig. 8). **c**, Fit of U2 SLIIa, BSL and SLI three-way junction into 17S U2 electron-microscopy density. **d**–**f**, Fit of the individual SLIIa (**d**), BSL (**e**) and a shortened SLI (**f**) into the electron-microscopy density. Owing to the formation of the BSL, only a shortened U2 SLI can form as nucleotides in the lower stem of an extended SLI would instead form base pairs located in the lower stem of the BSL. Thus, an extended SLI and a BSL are mutually exclusive, competing U2 snRNA conformations. The position of the remaining 10 nucleotides at the 5′ end of the U2 snRNA, which in the spliceosome form part of U2–U6 helix II, cannot be discerned. **g**, Fit of the SF3A3 separator helix into density at the base of the BSL. **h**, Similar SLIIa RNP architecture in 17S U2 snRNP, and the yeast B and human B^{act} spliceosomal complexes. In human U2 snRNP, loop nucleotides of SLIIa contact amino acids of the loop connecting the two α-helices of SF3B1 HR20, as well as residues of SF3B2; two regions of the latter (amino acids 458–530 and 565–598) are located in well-resolved density close to SLIIa and SLI, respectively. These SLIIa contacts are similar to those found in yeast B and human B^{act} spliceosomes. Thus, they are a major, direct anchor point for SF3b on the U2 snRNA. The poor resolution of the cryo-EM structure of the human B complex in this region does not allow for an accurate comparison.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Fit of PRP5 and TAT-SF1 and their interaction with the SF3B1 HEAT domain. a, Schematic of the domain organization of human PRP5 (left) and TAT-SF1 (right), with amino acid boundaries of each domain indicated below. **b**, Fit of the PRP5 RecA1 and RecA2 domains in an open conformation into the 17S U2 electron-microscopy density. The open (inactive) conformation of the PRP5 RecA domains seems to fit better than the closed (active) conformation, which suggests that PRP5 is inactive in 17S U2 snRNP. However, neither the resolution in this region nor intramolecular PRP5 crosslinks allow us to confidently position these two domains relative to each other and thereby distinguish between these two conformations. DEAD-box proteins typically bind double-stranded RNA via their RecA domains and facilitate local strand separation by introducing one or two sharp bends in one of the bound strands that prevent base-pairing with the complementary strand[28]. The resolution of the 17S U2 in the region where the PRP5 RecA domains are located does not allow us to discern whether PRP5 is interacting with RNA at this stage. On the basis of our structure, the RecA domains of PRP5 in 17S U2 are not close enough to the BSL to directly disrupt it, and also are not located close enough to the SLI. **c**, Protein crosslinks support the positioning of the PRP5 RecA domains and the TAT-SF1$^{\mathrm{UHM}}$. Crosslinks are observed between TAT-SF1$^{\mathrm{UHM}}$ and residues in the N- and C-terminal HEAT repeats of SF3B1, in the region of PRP5 between the N-terminal α-helix and RecA domains, and in PHF5A, which supports the location of TAT-SF1$^{\mathrm{UHM}}$ in the less-well resolved density element adjacent to the U2 BSL. **d**, Fit of the PRP5 α-helix (amino acids 146–196) into the electron-microscopy density contacting HEAT repeats 9–12. **e**, Protein crosslinks between PRP5 and SF3B1 suggest that a region spanning approximately 200 amino acids located N-terminal of the PRP5 helicase domain, wraps around most of SF3B1$^{\mathrm{HEAT}}$. These data are consistent with studies showing that yeast Prp5 also interacts with HR1–HR6 and HR9–HR12 of the yeast SF3B1 homologue, Hsh155[23]. Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The proposed path of unstructured regions of PRP5 is indicated by a dotted line and the location of the conserved PRP5 DPLD motif is shown. The position of selected cancer-related hotspot mutations of SF3B1 (K666, K662, K700 and G742) are indicated. Point mutations in SF3B1$^{\mathrm{HEAT}}$ are linked to various cancers and lead to the utilization of cryptic branch sites and 3′ splice sites in vivo[49,50]. The exact mechanism responsible for these changes in alternative splicing is currently unknown but it has been suggested that these mutations may affect the curvature of the HEAT domain[10]. Studies in yeast showed that position-equivalent, point mutations in SF3B1 that are linked to cancer in humans, lead to loss of stable Prp5 binding to the U2 snRNP[23,25]. More recent studies indicate that in humans these cancer-related mutations destabilize the interaction of SF3B1 with the SUGP1 protein[51], which, however, is essentially absent from our 17S U2 preparations[20]. As PRP5 appears to encompass the entire HEAT domain in the human 17S U2, it is conceivable that a change in the curvature of the HEAT domain could destabilize the SF3B1–PRP5 interaction, and the absence of PRP5 may directly lead to alterations in BS selection by U2 containing SF3B1 cancer-related point mutations. Prp5 was shown to contact the SF3b complex in the U2 snRNP via a conserved DPLD motif in its N-terminal domain[24]. The most common (hotspot) cancer-related point mutations mainly cluster in or near HR6[10] and notably our crosslinking data place the DPLD motif of PRP5 (amino acids 226–229) in proximity to HR6. Thus cancer-related hotspot mutations may disrupt this essential interaction, leading to destabilization of PRP5. This, in turn, could have a detrimental effect on the function of PRP5 to facilitate the formation of a stable U2–BS interaction and/or on the proofreading activity of PRP5 (as discussed above), and thus facilitate the usage of aberrant BS and 3′ splice sites. **f**, Protein crosslinks between TAT-SF1$^{\mathrm{RRM1}}$ and neighbouring proteins. In 17S U2, TAT-SF1$^{\mathrm{RRM1}}$ is located adjacent to SF3B1 HR15 and HR16 (Fig. 3a), and thus may also stabilize HR16, which is completely structured in 17S U2 (Extended Data Fig. 4c). Crosslinking also suggests that an unstructured loop of SF3B2 consisting of amino acids 531–564 may occupy the electron-microscopy density near SF3B1 HR12–HR15. **g**, Position of SF3B1 and SF3B2 in 17S U2 (left) and in the human $\mathrm{B}^{\mathrm{act}}$ complex (right). In 17S U2, TAT-SF1$^{\mathrm{RRM1}}$ is located in the same position where two α-helices of SF3B2 are found in the $\mathrm{B}^{\mathrm{act}}$ complex, and thus release of TAT-SF1 would be required for the subsequent formation or repositioning of this region of SF3B2. **h**, Fit of TAT-SF1 UHM (amino acids 260–353) into the 17S U2 electron-microscopy density. Amino acids 260–353 of human TAT-SF1 were initially designated RRM2, but were later shown to comprise a UHM[52].



Extended Data Fig. 8 | RNP rearrangements and movements of the U2 snRNA required for formation of the extended U2–BS helix. a, Spatial orientation of the BSL and neighbouring proteins in the 17S U2 (left), and the subsequently rearranged U2 snRNA and U2 proteins after formation of the U2/BS helix and stable U2 incorporation into the spliceosome (right). Although the human B^{act} structure (PDB code 6FF4) is shown, the spatial organization of the shown U2 components is similar in B and presumably also human A complexes. **b**, Movement of U2 SLI behind the BSL requires repositioning of the latter. Coloured surfaces are derived from the fitted protein models, with the RNA depicted as a combination of a transparent surface model and RNA helix. The movement of the 5′ end of U2 required for formation of the extended U2–BS helix would require release of the BSL from SF3B1. In the absence of the latter, the 5′ end of U2 snRNA, including the short SLI, which is topologically located above the BSL stem, would have to be threaded through a very narrow opening between the BSL and the SF3B1 HEAT domain in order to unwind the BSL, a scenario that is unlikely. Thus, the remodelling of the U2 BSL into an extended U2–BS helix will probably occur in a conformational state of the U2 snRNP in which the BSL is moved away from the C-terminal HEAT repeats. This movement could involve a rotation around U2 nucleotides U46 and/or U47, which link the BSL to the stem of SLIIa and appear to be maintained in a single-stranded conformation by SF3A3 (Extended Data Fig. 6). In this respect, it is intriguing that the short SF3A3 separator helix is situated at the same place in 17S U2 and in human B^{act} complexes. Moreover, it probably has very similar roles in 17S U2 and the spliceosome. That is, in B^{act}, the SF3A3 separator helix also determines the length of the extended part of the U2–BS helix and facilitates the movement of U2 snRNA and the intron away from each other. The SF3A3 separator helix probably maintains its contact with the U2 SLIIa, ensuring the stable interaction of the SF3b complex with the U2 snRNA. This has the advantage that the newly formed extended U2–BS complex can swing back towards the HEAT domain with SF3A3 still bound to the SLIIa. Docking of the extended U2–BS helix to the SF3B1 C-terminal HEAT repeats is potentially facilitated by initial interactions with the backbone of the first 5–6 nucleotides of the intron downstream of the BS. **c**, Movements of the 5′ end of U2 that enable formation of the extended U2–BS helix. Step 1: ATP hydrolysis by PRP5 disrupts protein contacts with the BSL and presumably also with the 5′ end of the U2 snRNA including SLI. This allows base pairing between the BSL loop nucleotides and the BS of the pre-mRNA intron. Step 2: the destabilized BSL unwinds and the 5′ end of U2 moves behind the BSL and downward, with a twisting motion that repositions the pre-mRNA behind the unwound BSL. Movement of the 5′ end downward allows the formation of additional base pairs with the pre-mRNA. As both ends of the pre-mRNA intron appear to be fixed by protein and snRNP interactions, formation of the helical conformation of the U2–BS and extended U2–BS probably involves first movement of the 5′ end of U2 behind the pre-mRNA (step 4), followed by a movement across the pre-mRNA (that is, a twisting rotation of the 5′ end of U2 around the pre-mRNA) (step 5). Numbers indicate the position of selected U2 nucleotides.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	17S U2 5' domain (EMD-10688) (PDB 6Y50)	17S U2 low resolution part (EMD-10689) (PDB 6Y53)	17S U2 particle (EMD-10689) (PDB 6Y5Q)
Data collection and processing			
Magnification	120,700	120,700	120,700
Voltage (kV)	300	300	300
Electron exposure (e$^-$/Å^2)	72	72	72
Defocus range (μm)	1-4	1-4	1-4
Pixel size (Å)	1.16	1.16	1.16
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	3.5 million	3.5 million	3.5 million
Final particle images (no.)	120,070	120,070	120,070
Map resolution (Å)	4.1	7.1	7.1
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)			
5'domain	3.6-9	---	4-9
TAT-SF1 RRM1	5-6	---	6-9
TAT-SF1 UHM	---	15	15
PRP5 RecA domains	---	20	20
3'domain	---	20	20
SF3A	---	30	30
Refinement			
Initial model used (PDB code)	n/a	n/a	n/a
Model resolution (Å)	4.2	---	---
FSC threshold	0.5		
Model resolution range (Å)	4.2	---	---
Map sharpening *B* factor (Å^2)	-241	-450	-450
Model composition			
Non-hydrogen atoms	14917	11306	22867
Protein residues	2537	2226	4763
Ligands	3	0	0
B factors (Å^2)			
Protein	54.74	---	---
Ligand	196.62	---	---
R.m.s. deviations			
Bond lengths (Å)	0.004	---	---
Bond angles (°)	1.18	---	---
Validation			
MolProbity score	2.05	---	---
Clashscore	13.78	---	---
Poor rotamers (%)	0	---	---
Ramachandran plot			
Favored (%)	93.94	---	---
Allowed (%)	5.90	---	---
Disallowed (%)	0.16	---	---

Extended Data Table 2 | Summary of modelled proteins and RNA in the human 17S U2 structure

Protein/RNA	Chain ID	UniProt ID	Total residues	Modeled Residue	Template	modeling approach
SF3A1	6	Q15459	793	160-286	4DGW	Docked
SF3A2	7	Q15428	464	104-209	4DGW	Docked
SF3A3	9	Q12874	501	1-362	4DGW	Docked
				392-499	6FF4	Docked and adjusted
SF3B1	u	O75533	1304	399-419	2FHO	Docked
				463-1304	6EN4	Docked and adjusted
SF3B2	8	Q13435	895	458-598	6FF4	Docked and adjusted
				607-658, 681-693	5LSB	Docked
SF3B3	v	Q15393	1217	1-1204	6EN4	Docked and adjusted
SF3B4	o	Q15427	424	11-89, 101-181	5LSB	Docked
SF3B5	x	Q9BWJ5	86	15-80	6EN4	Docked and adjusted
PHF5A	y	Q7RTV0	110	6-98	6EN4	Docked and adjusted
SF3B6/p14	z	Q9Y3B4	125	17-93	2FHO	Docked
A’	a	P09661	255	2-163	5MQF	Docked
B”	b	P08579	225	6-98	5MQF	Docked
				149-225	6FF4	Docked
Sm D2	h	P62316	118	19-116	5MQF	Docked
Sm F	i	P62306	86	2-75	5MQF	Docked
Sm E	j	P62304	92	14-92	5MQF	Docked
Sm G	k	P62308	76	3-76	5MQF	Docked
Sm D3	l	P62318	126	2-84	5MQF	Docked
Sm B	m	P14678	240	4-49,63-87	5MQF	Docked
Sm D1	n	P62314	119	1-82	5MQF	Docked
PRP5	p	Q7L014	1031	146-195		Predicted model, Docked
				350-743	4LJY	Docked and adjusted
TAT-SF1	q	O43719	755	127-220		Predicted model, docked and adjusted
				260-347	2DIT	Docked
U2snRNA	2	NR_002716	188	12-14,19-21		de novo, Docked
				22-24		de novo
				25-45		Predicted model, Docked
				46-47		de novo
				48-65	5GM6	Docked
				69-73, 81-85		de novo, Docked
				97-184	5MQF	Docked

RNA and protein regions were modelled and fit into the electron-microscopy density as indicated.

nature research

Corresponding author(s): Holger Stark

Last updated by author(s): March 3, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo Fischer EPU 2.1, Thermo Exactive MS Series ICSW v.2.2
Data analysis	MotionCor2, Gctf v1.06 , COW-MicrographQualityChecker, Gautomatch v0.56, RELION 3.0, UCSF Chimera v.1.13.1, cryoSPARC v2.1, Coot v. 0.8.9.2, SWISS-MODEL suite, SpliProt3D, pLink 1.23, pLink 2.3.5

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The coordinate files have been deposited in the Protein Data Bank as follows: U2 5' domain (6Y50), low resolution region (6Y53) and entire 17S U2 particle (6Y5Q). The cryo-EM maps have been deposited in the Electron Microscopy Data Bank as follows: U2 5' domain (EMD-10688) and entire U2 particle (EMD-10689).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size does not apply to this study because it deals with a macromolecular structure.
Data exclusions	No data were excluded
Replication	Experimental replication was not attempted because this study deals with a macromolecular structure.
Randomization	Randomization does not apply because this study deals with a macromolecular structure.
Blinding	Blinding does not apply because this study deals with a macromolecular structure.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies against human SF3B1 were generated in house
Validation	Anti-human SF3B1 rabbit and mouse antibodies were validated in house for use in Western blotting and for immunoaffinity purification. See Will et al., (2002) EMBO J. 21:4978-88.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa S3 cells were obtained from GBF, Brunswick (currently Helmholtz Center for Infection Research)
Authentication	The HeLa S3 cell line was not authenticated.
Mycoplasma contamination	HeLa S3 cells tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Femtosecond-to-millisecond structural changes in a light-driven sodium pump

https://doi.org/10.1038/s41586-020-2307-8

Received: 29 November 2019

Accepted: 16 April 2020

Petr Skopintsev[1], David Ehrenberg[2,13], Tobias Weinert[1,13], Daniel James[1], Rajiv K. Kar[3], Philip J. M. Johnson[4], Dmitry Ozerov[5], Antonia Furrer[1], Isabelle Martiel[6], Florian Dworkowski[6], Karol Nass[6,7], Gregor Knopp[7], Claudio Cirelli[7], Christopher Arrell[8], Dardan Gashi[1,7], Sandra Mous[9], Maximilian Wranik[1], Thomas Gruhl[1], Demet Kekilli[1], Steffen Brünle[1], Xavier Deupi[1,10], Gebhard F. X. Schertler[1,11], Roger M. Benoit[1,12], Valerie Panneels[1], Przemyslaw Nogly[9], Igor Schapiro[3], Christopher Milne[7], Joachim Heberle[2] & Jörg Standfuss[1]✉

Light-driven sodium pumps actively transport small cations across cellular membranes[1]. These pumps are used by microorganisms to convert light into membrane potential and have become useful optogenetic tools with applications in neuroscience. Although the resting state structures of the prototypical sodium pump *Krokinobacter eikastus* rhodopsin 2 (KR2) have been solved[2,3], it is unclear how structural alterations over time allow sodium to be translocated against a concentration gradient. Here, using the Swiss X-ray Free Electron Laser[4], we have collected serial crystallographic data at ten pump–probe delays from femtoseconds to milliseconds. High-resolution structural snapshots throughout the KR2 photocycle show how retinal isomerization is completed on the femtosecond timescale and changes the local structure of the binding pocket in the early nanoseconds. Subsequent rearrangements and deprotonation of the retinal Schiff base open an electrostatic gate in microseconds. Structural and spectroscopic data, in combination with quantum chemical calculations, indicate that a sodium ion binds transiently close to the retinal within one millisecond. In the last structural intermediate, at 20 milliseconds after activation, we identified a potential second sodium-binding site close to the extracellular exit. These results provide direct molecular insight into the dynamics of active cation transport across biological membranes.

The preservation of sodium gradients across cellular membranes is crucial for various biological functions. In living cells, the controlled flow of sodium is maintained by a series of specialized membrane channels and pumps. For example, glucose uptake in the guts and kidneys of mammals is fuelled by a sodium gradient, which makes glucose–sodium symporters important pharmacological targets in the treatment of diabetes[5]. The opening and closing of voltage-gated sodium channels is responsible for the generation and propagation of neuronal signals. This has enabled the field of optogenetics, in which light-sensitive microbial cation channels from the rhodopsin family are used as a key component for the manipulation of physiological responses in neurons or even in living animals by light[6].

Rhodopsins are a functionally diverse family of proteins in microorganisms[7] and higher organisms[8] that rely on a retinal chromophore to harvest and sense light energy. In 2013, the rhodopsin family was extended by the discovery of light-driven sodium pumps in marine bacteria, where they maintain a low intracellular sodium ion concentration and generate a membrane potential[1]. In optogenetic applications, the controlled light-induced outward pumping of sodium ions leads to neuronal inhibition under more physiological conditions compared to the use of related proton or chloride pumps[2]. The optogenetic application of the prototypical member of the class, *Krokinobacter eikastus* rhodopsin 2 (KR2), has been demonstrated in nematodes and cortical rat neural cells[2]. Genetically engineered variants provide further possibilities to optimize KR2 for optogenetic applications[1–3,9–12].

The pumping cycle of KR2 has been studied using several time-resolved spectroscopic techniques[1,13–19]. High-resolution structures of the resting state have been determined in various forms by X-ray crystallography[2,3,20]. However, in these studies the sodium substrate is not bound within the retinal binding pocket, indicating that the pumping mechanisms is substantially different from those in related ion pumps. Additional structural information on the

---

[1]Laboratory of Biomolecular Research, Biology and Chemistry Division, Paul Scherrer Institut, Villigen, Switzerland. [2]Experimental Molecular Biophysics, Department of Physics, Freie Universität Berlin, Berlin, Germany. [3]Fritz Haber Center for Molecular Dynamics, Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel. [4]Laboratory of Non-linear Optics, Photon Science Division, Paul Scherrer Institut, Villigen, Switzerland. [5]Science IT, Paul Scherrer Institut, Villigen, Switzerland. [6]Laboratory for Macromolecules and Bioimaging, Photon Science Division, Paul Scherrer Institut, Villigen, Switzerland. [7]Laboratory of Femtochemistry, Photon Science Division, Paul Scherrer Institut, Villigen, Switzerland. [8]Laboratory for Advanced Photonics, Photon Science Division, Paul Scherrer Institut, Villigen, Switzerland. [9]Institute of Molecular Biology and Biophysics, Department of Biology, ETH Zürich, Zürich, Switzerland. [10]Condensed Matter Theory Group, Paul Scherrer Institut, Villigen, Switzerland. [11]Department of Biology, ETH Zürich, Zürich, Switzerland. [12]Laboratory of Nanoscale Biology, Division of Biology and Chemistry, Paul Scherrer Institut, Villigen, Switzerland. [13]These authors contributed equally: David Ehrenberg, Tobias Weinert. ✉e-mail: joerg.standfuss@psi.ch

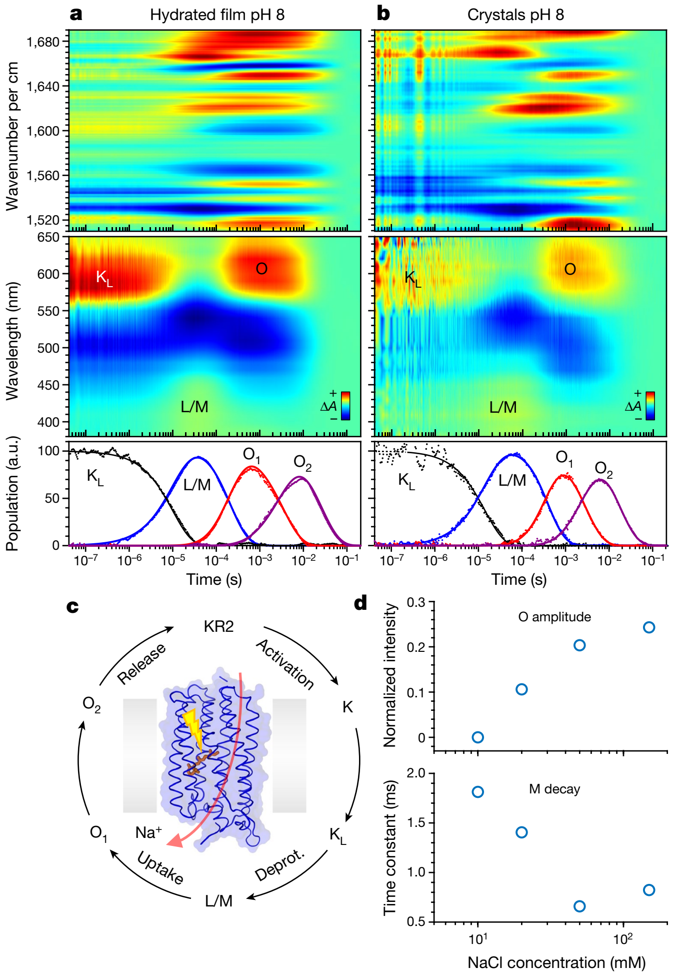


Fig. 1 | Time-resolved absorption measurements on KR2 in solution and crystals. a, **b**, Spectra from purified KR2 (**a**) and in the crystalline phase (**b**) prepared in analogy to the TR-SFX experiment. Top, experiments in the infrared covering the C=C stretch mode of the retinal up to amide I vibrations originating from the protein backbone. Middle, changes in the UV/Vis region on the same sample. Bottom, a global fit analysis of the infrared data reveals the presence of intermediate states K_L, L/M, O_1 and O_2. **c**, Model of the photocycle of KR2 derived from time-resolved absorption spectroscopy. **d**, Sodium dependency of the 1,516 cm^{-1} marker band for the O intermediate in KR2 crystals under TR-SFX conditions (top, amplitude of O intermediate; bottom, decay of the preceding M intermediate). Time traces were normalized to the ground state bleach signal and fitted by the sum of three connected exponentials.

reaction intermediates is required to understand how sodium can be actively transported out of the cell against substantial concentration gradients.

Time-resolved serial femtosecond crystallography (TR-SFX) can provide structural snapshots of proteins that can be assembled into molecular movies of protein function. A series of classical targets have been characterized using this method, including myoglobin[21], photoactive yellow protein[22,23] and bacteriorhodopsin (bR)[24–27]. Here, we used the Swiss X-ray Free Electron Laser (SwissFEL) to study structural changes in KR2 within a wide temporal window from 800 fs to 20 ms after activation. Ten structural snapshots positioned at temporal delays coinciding with the accumulation of intermediates identified by time-resolved absorption spectroscopy on crystals show how the energy captured by the retinal leads to structural rearrangements. Structural, spectroscopic and quantum chemical data indicate that sodium binds between N112 and D251 and is released via a second binding site between E11, N106 and E160 on the extracellular side. Our integrated work thus elucidates the structural changes associated with active transport of sodium ions across a biological membrane.

Photocycle in KR2 crystals

As a benchmark for the study of KR2 activation in the crystalline environment, we used time-resolved (TR) absorption spectroscopy in the infrared (IR) and ultraviolet/visible (UV/Vis) regions. As under the original acidic crystallization conditions KR2 exhibits an accelerated photocycle, we developed a soaking protocol to increase its pH in the presence of sodium ions (Extended Data Fig. 1). The treatment changes the colour of crystals from blue to red with associated changes in the retinal binding pocket. Most importantly in terms of function, KR2 in treated crystals follows a photocycle (Fig. 1) identical to that observed by our absorption spectroscopy on purified KR2, in agreement with previous reports[18,19].

Sodium is expected to bind to KR2 after deprotonation of the retinal Schiff base (SB) in the M intermediate, followed by release in the late O intermediate, because the transition between these spectroscopic intermediates is dependent on sodium concentration[28]. At acidic pH, characteristic O-related bands are absent (Extended Data Fig. 2), whereas at a higher pH in the presence of sodium these bands reach maximal amplitude a few milliseconds after activation. The TR-IR data show that the amplitude and kinetics of the M–O transition depend on sodium concentration (Fig. 1d). Clearly, KR2 in treated crystals responds to the presence of sodium with a kinetic profile compatible with light-driven sodium pumping. Accordingly, all crystals for the dynamic measurements described below were rebuffered in the presence of sodium before injection across the X-ray laser pulses.

Structural changes over time

In three days of beamtime during the first user run of SwissFEL, we collected 158,832 dark (pump laser off) and 496,904 light (pump laser on) indexable diffraction patterns with an anisotropic resolution up to 1.6 Å (Extended Data Table 1, Extended Data Fig. 3a). The light data were distributed over 10 time delays (Δt = 800 fs, 2 ps, 100 ps, 1 ns, 16 ns, 1 µs, 30 µs, 150 µs, 1 ms and 20 ms) between the optical pump and the X-ray probe pulses (Extended Data Fig. 4a). These time delays were selected on the basis of our TR-IR data and of previous ultrafast stimulated Raman spectroscopic experiments[16], to cover critical steps in the KR2 photocycle. The serial crystallographic structure of the KR2 resting state closely resembles structures solved by conventional cryo-crystallography (Extended Data Fig. 5). Progressing from this starting point, isomorphous difference electron density maps (F_o(light) − F_o(dark)) allowed us to follow structural changes over time. Extrapolated data were used to refine the molecular structures for each time delay. The light-activated structures follow a continuous evolution of structural rearrangements that we combined into five stages (800 fs + 2 ps, 1 ns + 16 ns, 30 µs + 150 µs, 1 ms and 20 ms) on the basis of root mean square deviations (r.m.s.d.) between the models (Extended Data Fig. 4b). These five structural intermediates provide direct molecular insights into the sequence of structural rearrangements during the KR2 pumping cycle.

Retinal and transmission of light energy

The first stage in the structural evolution of KR2 activation ranges from femtoseconds to picoseconds, with structural rearrangements centred at the retinal chromophore, the principal photochemical switch of all rhodopsins. Retinal is covalently bound via a protonated SB linkage to K255 in helix G in the core of KR2. Similar to other microbial rhodopsins, absorption of a photon in KR2 leads to retinal isomerization at the C13=C14 bond (Extended Data Fig. 3b). This photochemical process is faster than in the

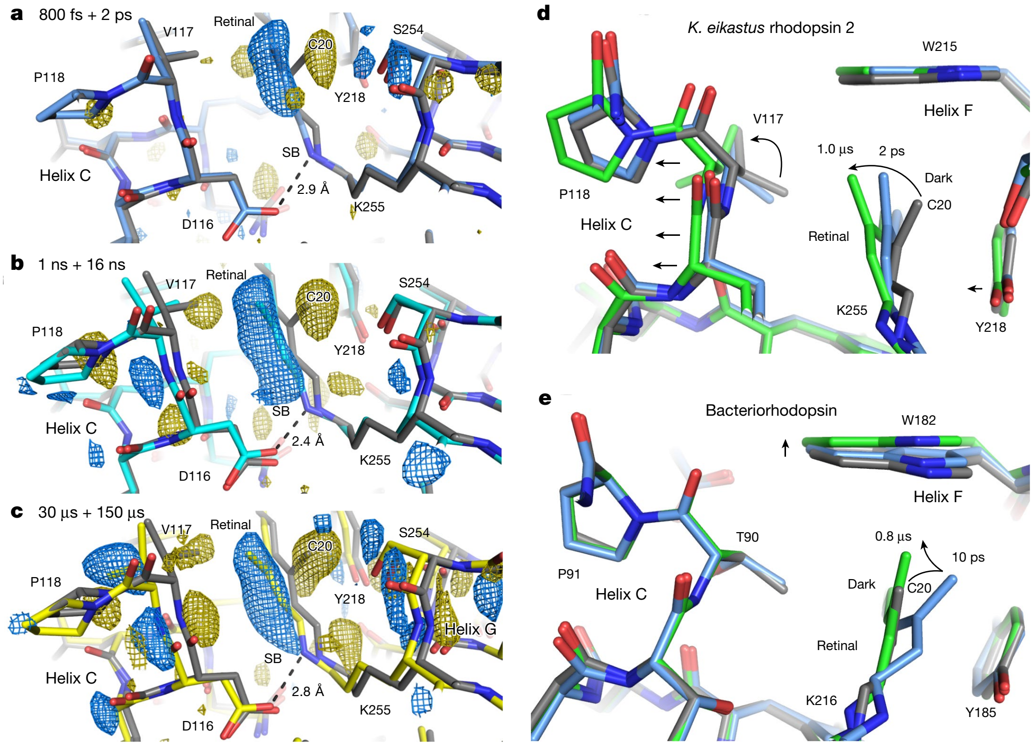


Fig. 2 | Early steps in the activation of a light-driven sodium pump. **a**–**c**, Difference electron density maps (F_o(light) − F_o(dark); gold, negative density; blue, positive density; shown at 3.5σ) and structural refinements (grey, resting state; colours, light-activated) show the early rearrangements at $\Delta t = 800$ fs + 2 ps (**a**), $\Delta t = 1$ ns + 16 ns (**b**) and $\Delta t = 30$ µs + 150 µs (**c**). **d**, **e**, Superposition of structures illustrates how retinal isomerization translates the light energy into structural changes via V117 in helix C in case of KR2 (**d**; grey sticks, resting state; blue sticks, structure at $\Delta t = 2$ ps; green sticks, structure at $\Delta t = 1$ µs) and via W182 in helix F in case of bR (**e**; data and coordinates previously published[24,25]; grey sticks, resting state; blue sticks, structure at $\Delta t = 10$ ps; green sticks, structure at $\Delta t = 0.8$ µs). Arrows indicate structural changes discussed in the main text.

related proton pump bR, with formation of the earliest photointermediate after only about 200 fs[16], which is consistent with a fully isomerized 13-*cis* retinal in our earliest difference electron density at $\Delta t = 800$ fs and after structural refinement at $\Delta t = 800$ fs + 2 ps (Fig. 2a). Within the early time delays, we further observed a shift of the water molecule w406 and the retinal counterion D116; this was similar to, but less pronounced than, the ultrafast adaptation of bR[25]. A clear difference from bR is that the isomerized retinal in KR2 points in the opposite direction, with the C20 methyl group tilting towards helix C instead of helix G. The direction could be pre-determined in the resting state as, compared to bR, the retinal polyene KR2 bends in the opposite direction.

The second distinct stage in the structural evolution takes place in the nanosecond range (Fig. 2b), when early conformational changes in the protein backbone occur. In bR, adaptation of the energetically more favourable planar 13-*cis* conformation pushes the straightening retinal 'upwards' against W182 in helix F to displace it towards the cytoplasmic side[24]. In KR2, the changes involving the structurally equivalent W215 are absent; instead, the retinal C20 methyl group pushes 'sideways' in the membrane plane against V117. Starting at $\Delta t = 1$ µs and rising until the later microsecond delay, the difference density maps indicate a flip of V117 and an established transmission of structural changes into helix C (Fig. 2c). In this way, the light energy stored in the early photoproducts propagates into the seven-transmembrane helical bundle to fuel larger conformational changes at later times.

Sodium translocation and gating

The third and fourth stages of the structural evolution from microseconds to milliseconds correlate with the temporal range that is relevant for sodium translocation[19,28]. Starting from $\Delta t = 1$ µs, clear electron difference density peaks above 3.5σ show how Y218 in helix F and S254 in helix G approach the position of retinal in the resting state. Changes are further transmitted several turns along helix G towards the intracellular side. Structural refinement resulted into small shifts of helix C in the order of 1 Å in the $\Delta t = 30$ µs + 150 µs structural intermediate, and additional changes along helix D occur in the 1-ms and 20-ms delays. These rearrangements (Fig. 3) are of particular interest because they are close to the putative entry and exit routes for sodium[2,3].

One half of the translocation pathway connects the retinal binding pocket between helices C, F and G with a water-filled cavity on the intracellular side. It passes N61 and G263 at the entry side and Q123 of the NQD motif (three residues that are important for ion selectivity[3,11,13]). Native KR2 pumps lithium and sodium ions, but not larger cations such as potassium[1]. In their hydrated forms, sodium is larger than potassium, but the dehydrated sodium ion is smaller than potassium[29,30]. This suggests that the selective pumping of sodium ions must involve dehydration, which is likely to happen at the entry to the conducting pathway[30].

The narrowest part of the sodium translocation pathway in the light-activated structures of KR2 passes between the retinal and the side chains of D116 and K255, which act as a counterion and a covalent link for the retinal, respectively. Neutralization of the retinal SB through proton transfer and transient widening at this position is likely to act as an electrostatic gate (Fig. 4a, Supplementary Video 1) that allows the selective passage of cations[2,3,31]. The distance between the retinal SB and D116 contracts by about 0.5 Å at $\Delta t = 1$ ns + 16 ns, which favours proton transfer in the transition to the M intermediate[32]. At $\Delta t = 30$ µs + 150 µs, the distance widens again in agreement with the spectroscopic data, indicating deprotonation of the SB and occurrence of the M state in the early microseconds (Fig. 1). At 1.4 Å, the opening seems small for a 1.9 Å sodium ion to pass. At this point we cannot exclude the possibility that the gate does not fully open in our crystals (which are formed from monomeric KR2), because pentameric KR2 can adopt a more open conformation in the SB region[3,20] and mutations in the oligomerization interface can affect pumping efficiency[1,17,20] (for further discussion, see Extended Data Fig. 5). However, stable structural intermediates that accumulate in time-resolved studies of molecule ensembles do not necessarily reveal all functional steps[33] and our integrated structural, spectroscopic and computational analysis is compatible with sodium binding in the later stages (see below). It is most likely that the approach of a sodium ion and electrostatic weakening of the helix C–SB interaction allows a transient pathway to form within the L–M equilibrium in the microseconds.

The fifth stage in the late milliseconds is related to changes in the section of the translocation pathway that extends from the retinal

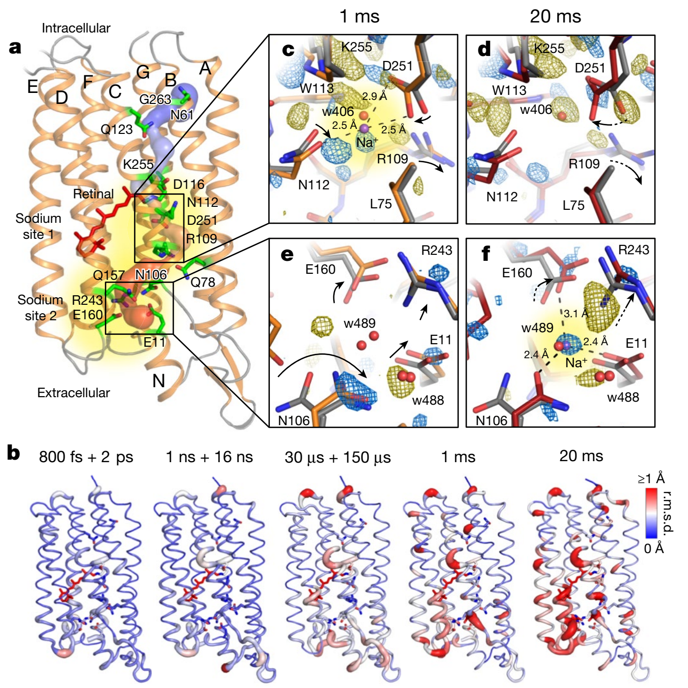


Fig. 3 | Structural changes along the sodium translocation pathway. **a**, Overview of the KR2 structure with the suggested route of sodium translocation across the membrane identified using the program Caver[39] (blue and red surfaces for the two halves plotted on the structure at $\Delta t = 1$ ms with changes in colour indicating functionally critical regions). Selected residues and the retinal chromophore are shown as sticks. **b**, Structural refinements and comparison to the resting state show how the protein rearranges from nanoseconds to milliseconds (blue-to-red gradient and ribbon width indicate r.m.s.d. to the resting state). **c**–**f**, Close-up views into the region of the retinal SB (**c**, **d**) and the extracellular side of the membrane (**e**, **f**). The structural rearrangements are compatible with the formation of a transient sodium binding site between N112 and D251 at $\Delta t = 1$ ms (**c**) and a second site further along the translocation pathway between E11, N106 and E160 at $\Delta t = 20$ ms after photoactivation (**f**). Difference electron density maps (F_o(light) − F_o(dark); gold, negative density; blue, positive density; countered at 3.2σ); arrows highlight structural changes discussed in the main text.

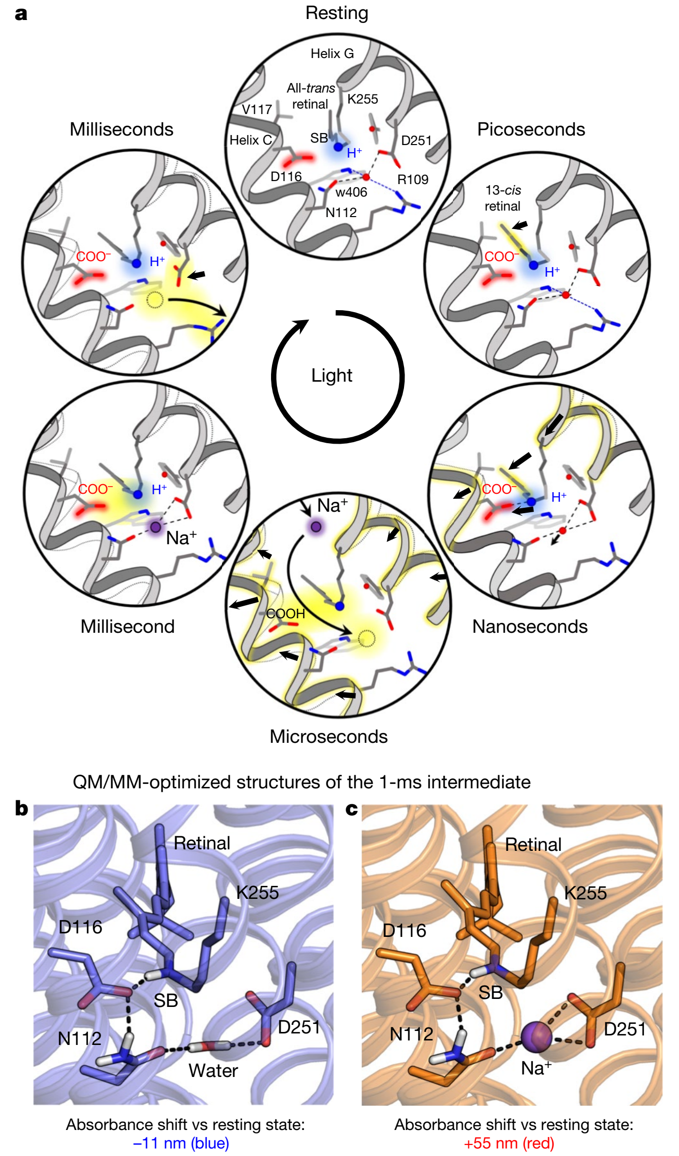


Fig. 4 | Electrostatic gating mechanism. **a**, Circles show the retinal binding pocket at critical steps in the KR2 photocycle. Blue, positive charge of the retinal SB; red, negative charge at the D116 counterion; arrows indicate steps in sodium translocation including light-induced retinal isomerization and conformational rearrangements in the retinal binding pocket, entry of a sodium ion after proton transfer from the SB to the counterion, binding of the sodium ion and reprotonation of the retinal SB and ion release. **b**, **c**, QM/MM-optimized geometry with water (**b**) or sodium ion (**c**) in the binding site between N112 and D251. The spectral shift is only in line with the spectroscopic data on the redshifted O intermediate (compare Fig. 1 and Extended Data Fig. 6h) when sodium is included. The shift is not a direct effect but is due to the change in the D116–SB interaction through N112 via the shown hydrogen-bonding network (dashed lines). For a dynamic illustration of the described rearrangements see Supplementary Video 1.

binding site towards the extracellular side of the protein. Here the bottlenecks run along the side chain of R109, consistent with mutations at this position transforming KR2 into a light-gated inward-facing potassium channel[34]. A rotamer change of R109 and Q78, together with a shift of helix D, at $\Delta t = 20$ ms indicates an opening that connects the water-filled cavity in the vicinity of the retinal with a second water-filled cavity close to E11, N106, E160 and R243, towards the exit site on the extracellular side of the membrane.

Formation of sodium binding sites

In the alternate access model of active membrane transport, the substrate is bound while the protein rearranges to allow release without backflow. Light-driven sodium pumps such as KR2 do not bind substrate close to the retinal SB in the resting state[1–3]. Hence, it remains unclear where the transported sodium ion is located and how it moves across the membrane with time. Time-resolved IR spectroscopy on a protein film (our results and published work[19]) and crystals provides evidence (through a marker band at 1,688 cm^{-1}) of changes in the environment of an asparagine residue that peak in the O intermediate within milliseconds under sodium pumping conditions (Fig. 1, Extended Data Fig. 2). Mutagenesis of N112 and D251 close to the retinal binding pocket abolish sodium pumping[13,34,35], and these two residues have been suggested to be potential sodium coordination partners, based on molecular dynamic simulations[31] and structural comparisons to other rhodopsins[36] (Extended Data Fig. 6). Analysis of our TR-SFX data shows a notable evolution of the electron density in this region (Fig. 3c, d). At $\Delta t = 1$ ms, a clear positive difference peak is located about 1 Å away from w406 in the resting state and close to N112 and D251. The electron densities of water and sodium ions have an identical signature in X-ray crystallographic data, but the environment can provide clues to the nature of the detected atoms[37]. At $\Delta t = 1$ ms, the density peak is shifted away from the amine group of W113 and the positively charged R109, both of which coordinate w406 in the resting state but cannot bind sodium. The new position is at a distance of 2.5 Å from N112 and D251, which is close to the ideal coordination distance for a sodium ion[38], concurring with our spectroscopic data. We validated the position of the putative sodium ion using hybrid quantum mechanics/molecular mechanics (QM/MM) calculations. Inclusion of the sodium in the $\Delta t = 1$ ms structure resulted in a large spectral red shift of 55 nm with respect to the electronic absorption band of the resting state, whereas placing water at this position yields an 11-nm blue shift (Fig. 4b, c). The experimental value obtained from the transient spectroscopic data is a red shift of 66 nm; hence, inclusion of the sodium ion is necessary to reproduce the absorption maximum of the O intermediate ($\lambda_{max} = 592$ nm). The absorption shift is not a direct effect, but results from changes in the D116–SB interaction through N112. It is reasonable to suggest that, through alteration of this H-bonding network, sodium binding favours reprotonation of the SB and thereby blocks the backflow of ions.

At $\Delta t = 20$ ms, the electron density close to the retinal binding pocket fades below the 3σ level, indicating release of the sodium ion. Further along the translocation pathway we observe the formation of a second sodium binding site close to the extracellular side of the membrane. Here, a clear positive difference peak appears between E11, N106 and E160. In the same temporal regime, the shift of R243 moves a positive charge away to facilitate sodium binding (Fig. 3e, f). Again, the coordination distances of 2.4 Å to N106 and 2.5 Å to E11 support our assignment of a sodium ion in this putative binding site. Notably, both sodium binding sites use displacements of arginine residues to favour binding of sodium over water. The corresponding positions of these residues have functional equivalents in bR (Extended Data Fig. 6), with R109 close to the retinal binding pocket corresponding to R82, which is critical for proton transfer in bR. The position of R243 in KR2 is occupied by E194 in bR, which is part of the proton release group. Some key sites in the seven-transmembrane helical bundle seem to be functionally conserved throughout evolution but are approached at different times within their respective pumping cycles.

Conclusions

Our data have allowed us to assemble a molecular movie of structural changes in KR2 and to propose a basic mechanistic model of light-driven sodium transport (Supplementary Video 1). The unidirectional flow of ions is achieved by minimal structural changes that generate ion selectivity and prevent ion back leakage into the cell. Our observation of active ion pumping is consistent with general concepts of ion pumping across a biological membrane by the alternate access model, illustrating them with high-resolution structures of the intermediate steps. Future studies will investigate how pH and long-range cooperative effects between protomers influence these structural dynamics. X-ray lasers now provide the means to study how single point mutations allow the translocation of larger ions such as potassium[3] and caesium[11] or turn KR2 from an active pump into a passive channel[34]. Deeper insights into the transport mechanisms found in microbial rhodopsins will demonstrate how evolution has adapted a common leitmotif to different functions and facilitate the design of variants for neurobiological applications in optogenetics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2307-8.

1. Inoue, K. et al. A light-driven sodium ion pump in marine bacteria. *Nat. Commun*. **4**, 1678 (2013).
2. Kato, H. E. et al. Structural basis for Na$^+$ transport mechanism by a light-driven Na$^+$ pump. *Nature* **521**, 48–53 (2015).
3. Gushchin, I. et al. Crystal structure of a light-driven sodium pump. *Nat. Struct. Mol. Biol*. **22**, 390–395 (2015).
4. Milne, C. et al. SwissFEL: The Swiss X-ray Free Electron Laser. *Appl. Sci. (Basel)* **7**, 720 (2017).
5. Poulsen, S. B., Fenton, R. A. & Rieg, T. Sodium–glucose cotransport. *Curr. Opin. Nephrol. Hypertens*. **24**, 463–469 (2015).
6. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci*. **18**, 1213–1225 (2015).
7. Govorunova, E. G., Sineshchekov, O. A., Li, H. & Spudich, J. L. Microbial rhodopsins: diversity, mechanisms, and optogenetic applications. *Annu. Rev. Biochem*. **86**, 845–872 (2017).
8. Terakita, A. The opsins. *Genome Biol*. **6**, 213 (2005).
9. Grimm, C., Silapetere, A., Vogt, A., Bernal Sierra, Y. A. & Hegemann, P. Electrical properties, substrate specificity and optogenetic potential of the engineered light-driven sodium pump eKR2. *Sci. Rep*. **8**, 9316 (2018).
10. Hoque, M. R. et al. A chimera Na$^+$-pump rhodopsin as an effective optogenetic silencer. *PLoS One* **11**, e0166820 (2016).
11. Konno, M. et al. Mutant of a light-driven sodium ion pump can transport cesium ions. *J. Phys. Chem. Lett*. **7**, 51–55 (2016).
12. Inoue, K. et al. Red-shifting mutation of light-driven sodium-pump rhodopsin. *Nat. Commun*. **10**, 1993 (2019).
13. Inoue, K., Konno, M., Abe-Yoshizumi, R. & Kandori, H. The role of the NDQ motif in sodium-pumping rhodopsins. *Angew. Chem. Int. Edn Engl*. **54**, 11536–11539 (2015).
14. Tahara, S. et al. Ultrafast photoreaction dynamics of a light-driven sodium-ion-pumping retinal protein from *Krokinobacter eikastus* revealed by femtosecond time-resolved absorption spectroscopy. *J. Phys. Chem. Lett*. **6**, 4481–4486 (2015).
15. Tahara, S. et al. Origin of the reactive and nonreactive excited states in the primary reaction of rhodopsins: pH dependence of femtosecond absorption of light-driven sodium ion pump rhodopsin KR2. *J. Phys. Chem. B* **122**, 4784–4792 (2018).
16. Hontani, Y. et al. The photochemistry of sodium ion pump rhodopsin observed by watermarked femto- to submillisecond stimulated Raman spectroscopy. *Phys. Chem. Chem. Phys*. **18**, 24729–24736 (2016).
17. Kaur, J. et al. Solid-state NMR analysis of the sodium pump *Krokinobacter* rhodopsin 2 and its H30A mutant. *J. Struct. Biol*. **206**, 55–65 (2019).
18. Chen, H.-F. et al. Time-resolved FTIR study of light-driven sodium pump rhodopsins. *Phys. Chem. Chem. Phys*. **20**, 17694–17704 (2018).
19. Asido, M. et al. Time-resolved IR spectroscopy reveals mechanistic details of ion transport in the sodium pump *Krokinobacter eikastus* rhodopsin 2. *Phys. Chem. Chem. Phys*. **21**, 4461–4471 (2019).
20. Kovalev, K. et al. Structure and mechanisms of sodium-pumping KR2 rhodopsin. *Sci. Adv*. **5**, eaav2671 (2019).
21. Barends, T. R. M. et al. Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* **350**, 445–450 (2015).
22. Tenboer, J. et al. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346**, 1242–1246 (2014).
23. Pande, K. et al. Femtosecond structural dynamics drives the *trans*/*cis* isomerization in photoactive yellow protein. *Science* **352**, 725–729 (2016).
24. Nango, E. et al. A three-dimensional movie of structural changes in bacteriorhodopsin. *Science* **354**, 1552–1557 (2016).
25. Nogly, P. et al. Retinal isomerization in bacteriorhodopsin captured by a femtosecond X-ray laser. *Science* **361**, eaat0094 (2018).
26. Weinert, T. et al. Proton uptake mechanism in bacteriorhodopsin captured by serial synchrotron crystallography. *Science* **365**, 61–65 (2019).
27. Nass Kovacs, G. et al. Three-dimensional view of ultrafast dynamics in photoexcited bacteriorhodopsin. *Nat. Commun*. **10**, 3177 (2019).
28. Kato, Y., Inoue, K. & Kandori, H. Kinetic analysis of H$^+$–Na$^+$ selectivity in a light-driven Na$^+$-pumping rhodopsin. *J. Phys. Chem. Lett*. **6**, 5111–5115 (2015).
29. Shannon, R. D. Revised effective ionic-radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
30. Kandori, H., Inoue, K. & Tsunoda, S. P. Light-driven sodium-pumping rhodopsin: a new concept of active transport. *Chem. Rev*. **118**, 10646–10658 (2018).
31. Suomivuori, C.-M., Gamiz-Hernandez, A. P., Sundholm, D. & Kaila, V. R. I. Energetics and dynamics of a light-driven sodium-pumping rhodopsin. *Proc. Natl Acad. Sci. USA* **114**, 7043–7048 (2017).
32. Nishimura, N., Mizuno, M., Kandori, H. & Mizutani, Y. Distortion and a strong hydrogen bond in the retinal chromophore enable sodium-ion transport by the sodium-ion pump KR2. *J. Phys. Chem. B* **123**, 3430–3440 (2019).
33. Ourmazd, A. Cryo-EM, XFELs and the structure conundrum in structural biology. *Nat. Methods* **16**, 941–944 (2019).
34. Vogt, A. et al. Engineered passive potassium conductance in the KR2 sodium pump. *Biophys. J*. **116**, 1941–1951 (2019).
35. Abe-Yoshizumi, R., Inoue, K., Kato, H. E., Nureki, O. & Kandori, H. Role of Asn112 in a light-driven sodium ion-pumping rhodopsin. *Biochemistry* **55**, 5790–5797 (2016).
36. Gushchin, I. et al. Structure of the light-driven sodium pump KR2 and its implications for optogenetics. *FEBS J*. **283**, 1232–1238 (2016).
37. Zheng, H. et al. CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr. D* **73**, 223–233 (2017).
38. Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L. & Minor, W. Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem*. **102**, 1765–1776 (2008).
39. Chovancova, E. et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLOS Comput. Biol*. **8**, e1002708 (2012).

Methods

Cloning, protein expression and purification

The KR2 construct with a TEV-cleavable C-terminal 6×His-tag was cloned into the pStaby1.2 vector (Delphi Genetics). Protein expression was performed in C41(DE3) *Escherichia coli* cells. The cells grew in shaking smooth flasks with Luria broth at 37 °C. Expression was induced by addition of 1 mM β-D-thiogalactopyranoside (IPTG) at an optical density at 600 nm (OD_{600}) of ~0.8. Following overnight expression at 37 °C in the presence of 10 μM all-*trans* retinal, the bacterial cultures were harvested by centrifugation at 5,000*g* for 15 min. The cell pellets were disrupted with Avestin EmulsiFlex-C3 homogenizer at 15,000 psi in lysis buffer (20 mM Tris pH 8.0, 5% glycerol, 0.5% Triton X-100, 5 μg/ml DNase I and cOmplete protease inhibitor tablets, Roche), and the membrane fraction was collected with ultracentrifugation at 90,000*g*. The membrane pellet was resuspended with IKA T 25 Ultra-Turrax disperser in solubilization buffer that contained 50 mM Tris pH 8.0, 300 mM NaCl, cOmplete protease inhibitors, 1.0% *n*-dodecyl β-D-maltoside (DDM, Anatrace), and 0.2% cholesteryl hemisuccinate (CHS, Anatrace), and stirred overnight at 4 °C. The overnight suspension was subjected to a second round of ultracentrifugation before the supernatant was applied to immobilized metal affinity chromatography (IMAC), and further washed with IMAC buffer (50 mM Tris pH 8.0, 150 mM NaCl, 100 mM imidazole, 0.02% DDM, 0.04% CHS). The bound protein was eluted by addition of 500 mM imidazole in IMAC buffer. TEV protease was added, and the KR2–TEV cleavage solution was sealed in an 8 kDa M_w cutoff dialysis membrane and dialysed against 50 mM Tris pH 8.0, 150 mM NaCl, 0.02% DDM, 0.04% CHS buffer overnight. The TEV-cleaved solution was re-applied to the IMAC column, and the flow-through was collected and concentrated with a centrifugal filter device (Millipore 100 kDa M_w cutoff). The concentrated protein sample was loaded onto a HiLoad Superdex 75 prep grade 16/600 column (GE Healthcare) equilibrated with SEC buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.05% DDM, 0.01% CHS). The elution profile was monitored at 280 nm and 530 nm with Shimadzu UV-2401PC spectrophotometer, and the purest fractions were concentrated to ~100 mg/ml, flash frozen in liquid nitrogen and stored at −80 °C until further crystallization.

Crystallization and TR-SFX sample preparation

Crystallization was carried out in lipidic cubic phase (LCP) using conditions similar to those described[2]. The purified protein buffer and monoolein (1-oleoyl-rac-glycerol, Nu-Chek prep) were thoroughly mixed in a 4:7 v/v ratio through coupled gas-tight Hamilton syringes. The formed LCP was extruded through Hamilton needles into plastic B-Braun Omnifix-F syringes loaded with precipitant (200 mM sodium acetate pH 4.4, 150 mM $MgCl_2$, 35% PEG 200). Crystallization occurred overnight in the dark at 20 °C and yielded plate-like blue KR2 crystals with dimensions of 10–30 × 10–25 × 1–3 μm^3 (for a size distribution, see Extended Data Fig. 1c).

Following formation of crystals, the precipitant solution was washed out by soaking the LCP in excess 150 mM NaCl, 35% PEG 200 solution, two times for 48 h in total. The washed phase, with light-blue unbuffered crystals, was further harvested into gastight Hamilton syringes in 60 μl fractions, and doped with 33 μl monoolein and 3.0 μl 50% PEG 1500 to form a stable jetting phase. Before data collection, the phase with crystals was mixed with LCP prepared from monoolein and 1 M Tris pH 9.0, 150 mM NaCl, 35% PEG 200 through a three-way syringe coupler[40]. The volumes of the mixed phases were picked such that the water fraction of the mesophase would contain 200 mM Tris, 150 mM NaCl, 35% PEG 200 and PEG 1500 and the blue KR2 crystals changed colour to red upon mixing (Extended Data Fig. 1). While the mixing was done with Tris solution at pH 9.0, the final pH of the preparation was close to pH 8 as confirmed by litmus paper. We attribute this shift by one pH unit and the shift in colour transition compared to KR2 solution to possible residual buffer trapped in LCP and/or the buffering capacity of monoolein molecules. The jetting stability of the mixed phase was confirmed before the XFEL experiment with a high-speed camera setup as described[40] (Supplementary Video 2).

Experimental setup and XFEL data collection

The TR-SFX data on KR2 crystals were collected in February 2019 in three days at the Alvra experimental station of SwissFEL. X-ray pulses with a photon energy of 12 keV and a pulse energy of 400 μJ at a repetition rate of 50 Hz were used for the experiment. On average 180 μJ (9×10^{10} X-ray photons) per pulse were deposited onto a 2.6 × 3.6 μm^2 spot (FWHM), focused by Kirkpatrick–Baez mirrors. The X-ray intensity was adjusted using solid attenuators to maximize diffraction signal without disrupting the sample injector flow or damaging the detector. To reduce X-ray scattering, the air in the sample chamber was pumped down to 100–200 mbar while being substituted with helium. To reduce the amount of data the Jungfrau 16M detector was run in 4M mode excluding the outer panels.

KR2 crystals were loaded into a high viscosity injector connected to an HPLC pump[41]. The crystals were extruded into the pump–probe interaction point through a 75-μm capillary at a flow rate of 3.35 μl per minute. In the interaction point, the probe XFEL beam intersected with a circularly polarized pump beam originating from an optical parametric amplifier producing laser pulses with 150 fs duration ($1/e^2$), 575 nm wavelength and 3 μJ total energy in a focal spot of 80 × 80 μm^2 beam ($1/e^2$), corresponding to a maximal laser fluence of 59 mJ/cm^2 and laser power density of 397 GW/cm^2. Approximating the dose per KR2 molecule with the Lambert–Beer law for an average 19 × 16 × 2 μm^3 sized crystal, we estimated that 8.3–3.5 photons per retinal are absorbed, depending on the crystal's orientation and the position of the individual chromophore within the optically dense crystal. However, the average photon dose is certainly lower because scattering and reflection on the extruded material further reduce the dose, with estimates ranging from 20%[27] to 90%[42]. Our previous best estimate was 80% for TR-SFX experiments on bR[25], which would reduce the calculated doses to 1.7–0.7 photons per retinal. Another point to consider is that the Sn←S1 excited-state absorption is in the 400–500 nm region[15]. As such, the excited state is unlikely to absorb a 575-nm photon, further minimizing the chance of multi-photon absorption.

To cover the KR2 photocycle time delays between the pump laser and the probing, XFEL pulses were chosen at Δt = 800 fs, 2 ps, 100 ps, 1 ns, 16 ns, 1 μs, 30 μs, 150 μs and 1 ms. An additional time delay at 20 ms was created by shifting the laser focus position and using the 50 Hz XFEL repetition rate to create a delay to the pump laser.

Every fifth pulse of the pump laser was blocked, so that a series of four light-activated and one dark diffraction pattern were collected in sequence. Roughly 50,000 light-activated patterns were collected for each time delay, a high-quality dark data set was obtained by merging patterns of the fifth pulse, and ~50,000 patterns with pump laser off. These laser off images were also acquired for comparison, to confirm that the dark data in each cycle was not illuminated. Finally, about 40,000 patterns were obtained with no laser activation for untreated crystals at acidic pH, to compare with crystals soaked as described above. For clarity, the applied data collection scheme is illustrated in Extended Data Fig. 4.

Data processing

All data were indexed, integrated and merged using Crystfel[43,44] version 0.8.0. Data were indexed using the xgandalf[45] and taketwo[46] algorithms. Data were integrated using the --rings option in indexamajig. Patterns were merged using partialator with the following options: --model = unity, --iterations = 3. No per-pattern resolution cutoff was applied. Data showed diffraction anisotropy of about 2.2 × 2.2 × 1.6 Å along a*, b* and c*. The general resolution cutoff of 1.6 Å was chosen after evaluating all uncut data sets using the staraniso server[47], showing a maximal resolution of 1.6 Å for most data sets (Extended Data Table 1). Diffraction intensities of each individual data set output from partialator were cut at 1.6 Å resolution and subjected to the staraniso server again. The staraniso server determines the resolution cutoff for each data set based on the signal-to-noise ratio in a given resolution shell and then truncates data along these directions; in the case of KR2 this resulted in data truncation very close to the given ellipsoid dimensions. It is currently not possible to generate a merged data set from serial crystallographic data using staraniso, so we generated data sets containing only reflections that were kept by staraniso and generated statistics from them using CrystFEL. Refinements for dark data collected at acidic and neutral pH were carried out to the full resolution range against data obtained after anisotropic truncation. Since data extrapolation lowers the data quality, we lowered the resolution cutoff for structural refinements of the combined data (800 fs + 2 ps, 1 ns + 16 ns and 30 µs + 150 µs) to 2.25 Å resolution and of the smaller data sets at 1 ms and 20 ms to 2.5 Å. The truncated data sets were deposited to the worldwide Protein Data Bank (wwPDB)[48] together with the structures (Extended Data Table 2) refined against original (resting acidic pH and resting neutral pH) and extrapolated data (800 fs + 2 ps, 1 ns + 16 ns, 30 µs + 150 µs, 1 ms and 20 ms).

Calculation of difference density maps

All F_o(light) − F_o(dark) difference maps were calculated using PHENIX[49] using the multi-scaling option excluding amplitudes smaller than 3σ and resolutions lower than 10 Å in the anisotropy corrected data and phases of the refined neutral resting state.

Structure determination and refinement of KR2 resting state

The structure of the KR2 neutral resting state was solved using molecular replacement with PDB accession 3X3C[2] as search model. The structure of the KR2 acidic resting state was refined directly using the neutral model as a starting point. Structural refinements were done using PHENIX[49] with iterative cycles of manual adjustments made in Coot[50].

Data extrapolation

Extrapolated data were calculated using the anisotropy corrected data and a linear approximation as follows: $F_{\mathrm{Extra}} = 100/A \times (F_o(\mathrm{light}) - F_o(\mathrm{dark})) + F_o(\mathrm{dark})$, where A is the activation level in per cent and F_{Extra} is the extrapolated structure factor amplitude. F_0(light) was scaled to F_o(dark) before calculation of F_{Extra}. The activation level A (the percentage of molecules that neither stayed in nor returned to the dark state after the laser pulse) was determined by calculating extrapolated maps with phases of the dark state and light data at different activation levels in steps of 2% in the calculation of F_{Extra}, until features of the dark state appeared at the retinal. On the basis of this analysis, we chose an activation level of 14% for extrapolated maps.

Refinement of time-resolved states

Negative amplitudes resulting from the extrapolation procedure were removed from the TR-SFX data before model building and refinement. The models were manually adjusted to best fit observed difference map features as well as extrapolated maps. For residues with multiple conformations present in the resting state, the prevalent conformation was chosen based on electron density maps, followed by refinement in PHENIX[49]. Initial models were refined against extrapolated data from all ten different time delays. Using a pairwise comparison at all time delays (Extended Data Fig. 4b) of structures based on their r.m.s.d., in combination with manual inspection of electron density maps and input from the time-resolved spectroscopic data (Fig. 1), we identified structural transition points and combined delays at 800 fs with 2 ps, 1 ns with 16 ns and 30 µs with 150 µs to further improve density maps and refinement statistics and model quality for the first three deposited intermediates (Extended Data Table 2). In the final refinement of the two delays in the millisecond regime, restraints on sodium distance were used. However, this only marginally affected results, as the sodium atom close to the retinal binding pocket refined to a distance of 2.4 Å to N112 and 2.7 Å to D251 with an average near identical to the 2.5 Å from restrained refinement. A similar result was obtained within the second binding pocket where the sodium atom refined to a distance of 2.4 Å to N106 and 2.5 Å to E11 with constrains and 2.4 Å to both residues without them.

Time-resolved spectroscopy

Crystals for spectroscopic characterization were prepared in a similar fashion to those used in the time-resolved SwissFEL experiments. A few microlitres of crystals immersed in LCP (lipidic cubic phase) were sandwiched between two BaF_2 windows and sealed with vacuum paste immediately after extrusion to prevent drying[51]. Similar to our previous studies[52,53], KR2 solubilized in 0.02% DDM, 0.004% CHS with 100 mM Tris pH 8.0, 150 mM NaCl, was dried from a concentrated protein solution on a BaF_2 window. The dried film was rehydrated via the vapour phase generated by a glycerol:water mixture in a 3:2 ratio and sealed with a second window using vacuum grease.

Time-resolved IR experiments were recorded on a home-built spectrometer based on tunable Quantum Cascade Lasers (QCLs) as described[54]. We traced transient absorption changes in the frequency range of 1,510–1,690 cm^{-1} in steps of 2 cm^{-1} for crystals and 1 cm^{-1} for protein films across the time range of 5 ns to 200 ms. The repetition rate was set to 2 Hz and each kinetic was averaged 25 times. After reaching 1,690 cm^{-1}, the scanning direction was reversed and the two data sets were merged, accounting for possible protein bleaching. The frequency-doubled emission of a Q-switched Nd:YAG (neodymium yttrium aluminium garnet, Minilite; Amplitude) laser emitting at 532 nm was used for photoactivation with an energy density set to ~3 mJ/cm^2.

Absorption changes in the visible range were recorded using a commercial flash photolysis setup (LKS70; Applied Photophysics) essentially as described[52]. The photoreaction was induced by a short laser pulse emitted by a Nd:YAG laser (Quanta-Ray; Spectra-Physics), which drives an optical parametric oscillator tuned to 532 nm with an energy density of ~3 mJ/cm^2. Transients were recorded from 380 to 650 nm in 10-nm steps omitting the wavelength around the exciting laser pulse owing to light scattering (that is, 510–550 nm). Each trace was recorded 10 times with a repetition rate of 2 Hz and subsequently averaged.

Time-resolved step-scan and rapid-scan FTIR experiments on KR2 films were conducted using a Vertex 80v spectrometer (Bruker Optics). The excitation laser source was the same as for the QCL measurements with a repetition rate of 2 Hz for the step-scan mode. In the case of rapid-scan experiments, a slower repetition rate of 0.018 Hz was used to take account of the slower photocycle of KR2 in the presence of KCl.

We reconstructed the recorded data by applying singular value decomposition. Kinetic analysis was done by fitting the data to a model consisting of a unidirectional sequence of states. This yields a concentration profile of the involved and spectroscopically observable states over the course of our measurement[53].

Hybrid QM/MM calculations

Crystallographic coordinates from TR-SFX were used as the initial input for the hybrid QM/MM calculations[55]. A subsystem of the protein was chosen and treated using a quantum chemical method (QM region), while the remaining part was treated using a classical force field (MM region), namely AMBER ff14SB[56]. The QM region includes the retinal chromophore, the side chain of K255, and side chains of residues D116, N112, and D251. The link atom was placed at the QM/MM boundary between the Cα and Cβ atoms of the amino acids. To test the effect of sodium binding we included either a sodium ion or a water in the QM region of the 1-ms structure.

The geometry optimization was performed at the B3LYP/cc-pVDZ/AMBER level of theory. In these simulations, the protein backbone position was fixed to the crystallographic structure, whereas the QM region and side chains of amino acids within the 5 Å region of the retinal protonated Schiff base in the MM region was relaxed. Resolution of identity for Coulomb integrals (RI-J) and chain of sphere approximation for the exchange integrals were applied (RIJCOSX)[57]. Correction for dispersion was included with D3/B-J damping variant[58]. The calculation of vertical excitation energies used the simplified TD-DFT method[59] developed by Grimme and co-workers at the CAM-B3LYP/cc-pVTZ level of theory. All the QM/MM calculations were performed with the quantum chemistry package Orca interfaced with the DL_POLY module of ChemShell[60] for the force field.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Resting state coordinates and structure factors have been deposited in the PDB database under accession code 6TK7 (acidic pH) and 6TK6 (neutral pH). Together with the neutral pH resting state structure, the light-activated data for all time points (800 fs, 2 ps, 100 ps, 1 ns, 16 ns, 1µs, 30 µs, 150 µs, 1ms and 20 ms) were deposited in the mmCIF file. For the refined structures using combined data (800 fs + 2 ps, 1 ns + 16 ns and 30 µs + 150 µs) and single (1 ms and 20 ms) light-activated data sets, coordinates, light amplitudes, dark amplitudes and extrapolated structure factors have been deposited in the PDB database under accession codes 6TK5 (800 fs + 2 ps), 6TK4 (1 ns + 16 ns), 6TK3 (30 µs + 150 µs), 6TK2 (1 ms) and 6TK1 (20 ms).

40. James, D. et al. Improving high viscosity extrusion of microcrystals for time-resolved serial femtosecond crystallography at X-ray lasers. *J. Vis. Exp.* **144**, e59087 (2019).
41. Weierstall, U. et al. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat. Commun.* **5**, 3309 (2014).
42. Claesson, E. et al. The primary structural photoresponse of phytochrome proteins captured by a femtosecond X-ray laser. *eLife* **9**, e53514 (2020).
43. White, T. A. et al. CrystFEL: a software suite for snapshot serial crystallography. *J. Appl. Crystallogr.* **45**, 335–341 (2012).
44. White, T. A. et al. Recent developments in CrystFEL. *J. Appl. Crystallogr.* **49**, 680–689 (2016).
45. Gevorkov, Y. et al. XGANDALF - extended gradient descent algorithm for lattice finding. *Acta Crystallogr. A* **75**, 694–704 (2019).
46. Ginn, H. M. et al. TakeTwo: an indexing algorithm suited to still images with known crystal parameters. *Acta Crystallogr. D* **72**, 956–965 (2016).
47. Tickle, I. J. et al. Staraniso. http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi (2020).
48. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
49. Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
50. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
51. Heberle, J., Büldt, G., Koglin, E., Rosenbusch, J. P. & Landau, E. M. Assessing the functionality of a membrane protein in a three-dimensional crystal. *J. Mol. Biol.* **281**, 587–592 (1998).
52. Lórenz-Fonfría, V. A. et al. Transient protonation changes in channelrhodopsin-2 and their relevance to channel gating. *Proc. Natl Acad. Sci. USA* **110**, E1273–E1281 (2013).
53. Ehrenberg, D. et al. The two-photon reversible reaction of the bistable jumping spider rhodopsin-1. *Biophys. J.* **116**, 1248–1258 (2019).
54. Schultz, B.-J., Mohrmann, H., Lórenz-Fonfría, V. A. & Heberle, J. Protein dynamics observed by tunable mid-IR quantum cascade lasers across the time range from 10 ns to 1 s. *Spectrochim. Acta A* **188**, 666–674 (2018).
55. Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Edn Engl.* **48**, 1198–1229 (2009).
56. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
57. Kossmann, S. & Neese, F. Efficient structure optimization with second-order many-body perturbation theory: the RIJCOSX-MP2 method. *J. Chem. Theory Comput.* **6**, 2325–2338 (2010).
58. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
59. Grimme, S. A simplified Tamm–Dancoff density functional approach for the electronic excitation spectra of very large molecules. *J. Chem. Phys.* **138**, 244104 (2013).
60. Metz, S., Kästner, J., Sokol, A., Keal, T. & Sherwood, P. ChemShell—a modular software package for QM/MM simulations. *WIREs Comput. Mol. Sci.* **4**, 101–110 (2014).
61. Balashov, S. P. et al. Light-driven Na^+ pump from *Gillisia limnaea*: a high-affinity Na^+ binding site is formed transiently in the photocycle. *Biochemistry* **53**, 7549–7561 (2014).
62. Noguchi, T., Ono, T.-A. & Inoue, Y. Direct detection of a carboxylate bridge between Mn and Ca^{2+} in the photosynthetic oxygen-evolving center by means of Fourier transform infrared spectroscopy. *Biochim. Biophys. Acta* **1228**, 189–200 (1995).
63. Nara, M. et al. Infrared studies of interaction between metal ions and Ca^{2+}-binding proteins. Marker bands for identifying the types of coordination of the side-chain COO– groups to metal ions in pike parvalbumin (pI = 4.10). *FEBS Lett.* **349**, 84–88 (1994).
64. Nara, M., Morii, H. & Tanokura, M. Coordination to divalent cations by calcium-binding proteins studied by FTIR spectroscopy. *Biochim. Biophys. Acta* **1828**, 2319–2327 (2013).
65. Furutani, Y., Murata, T. & Kandori, H. Sodium or lithium ion-binding-induced structural changes in the K-ring of V-ATPase from *Enterococcus hirae* revealed by ATR-FTIR spectroscopy. *J. Am. Chem. Soc.* **133**, 2860–2863 (2011).
66. Wickstrand, C. et al. A tool for visualizing protein motions in time-resolved crystallography. *Structural Dynamics (Melville, N.Y.)* **7**, 024701 (2020).
67. Kim, K. et al. Crystal structure and functional characterization of a light-driven chloride pump having an NTQ motif. *Nat. Commun.* **7**, 12677 (2016).
68. Kolbe, M., Besir, H., Essen, L. O. & Oesterhelt, D. Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science* **288**, 1390–1396 (2000).
69. Wickstrand, C. et al. Bacteriorhodopsin: structural insights revealed using X-ray lasers and synchrotron radiation. *Annu. Rev. Biochem.* **88**, 59–83 (2019).
70. Gerwert, K., Freier, E. & Wolf, S. The role of protein-bound water molecules in microbial rhodopsins. *Biochim. Biophys. Acta* **1837**, 606–613 (2014).

Acknowledgements This project was funded by the following agencies: The Swiss National Science Foundation project grants 31003A_141235 and 31003A_159558 (to J.S.), PZ00P3_174169 (to P.N.), 310030_192780 (to X.D.) and 310030B_173335 (to G.F.X.S.). We further acknowledge support by the NCCR:MUST (to C.M. and J.S.). The German Research Foundation supported the work through SFB-1078, project B3 and SPP-1926, HE 2063/6-1 (to J.H). This project has received funding from the European Union's Horizon 2020 research and innovation program under Marie-Sklodowska-Curie grant agreements 701646 and 701647. I.S. acknowledges funding by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant no. 678169 'PhotoMutant'). I.S. thanks the SFB 1078 'Protonation Dynamics in Protein Function' for the Mercator fellowship. R.K.K. acknowledges support from the Lady Davis Trust for the Arskin postdoctoral fellowship. We thank N. Varma for discussions on TR-SFX sample jetting and the Macromolecular Crystallography group for support during testing of crystals at the Swiss Light Source. We further thank everybody involved in ensuring the smooth operation of the Swiss X-ray Free Electron Laser during our experiments.

Author contributions The project was initiated and coordinated by J.S. following input on crystallographic data evaluation from T.W., quantum chemical calculations from I.S., pump–probe experiments at the Alvra endstation from C.M. and time-resolved spectroscopy from J.H. G.F.X.S. coordinated and supported crystallographic applications at SwissFEL and contributed to discussions throughout the project. The KR2 expression construct was cloned by R.M.B. The protein was expressed, prepared and crystallized by P.S. with help from A.F. A constant supply of sample during the beamtime was secured with the help of A.F., T.G., M.W., D.E., D.K. and R.M.B. using protocols developed by P.S. Crystal injection was optimized by P.S. and D.J. The lipidic cubic phase injector was operated and aligned during the beamtime by D.J., F.D., I.M. and D.G. The endstation including the laser system was aligned and operated by P.S., K.N., G.K., C.C. and C.M., who also designed the Alvra prime pump–probe station. C.A. and P.J.M.J. operated the timing diagnostics during the experiment. The SFX data analysis pipeline was built and operated by D.O. and K.N. Data processing during the beamtime was done by T.W., P.N. and S.B. S.M., V.P. and X.D. recorded progress during data collection and/or in a previous test beamtime. Final structures were refined by P.S. and T.W. and interpreted together with J.S. Quantum chemical calculations were done by R.K.K. and I.S. The time-resolved spectroscopic experiments were done by D.E. and P.S. and interpreted together with J.H.. The manuscript was written by J.S. with direct contributions from P.S., T.W., D.E., J.H., R.K.K. and I.S. and with further suggestions from most of the other authors. All authors read and acknowledged the manuscript.

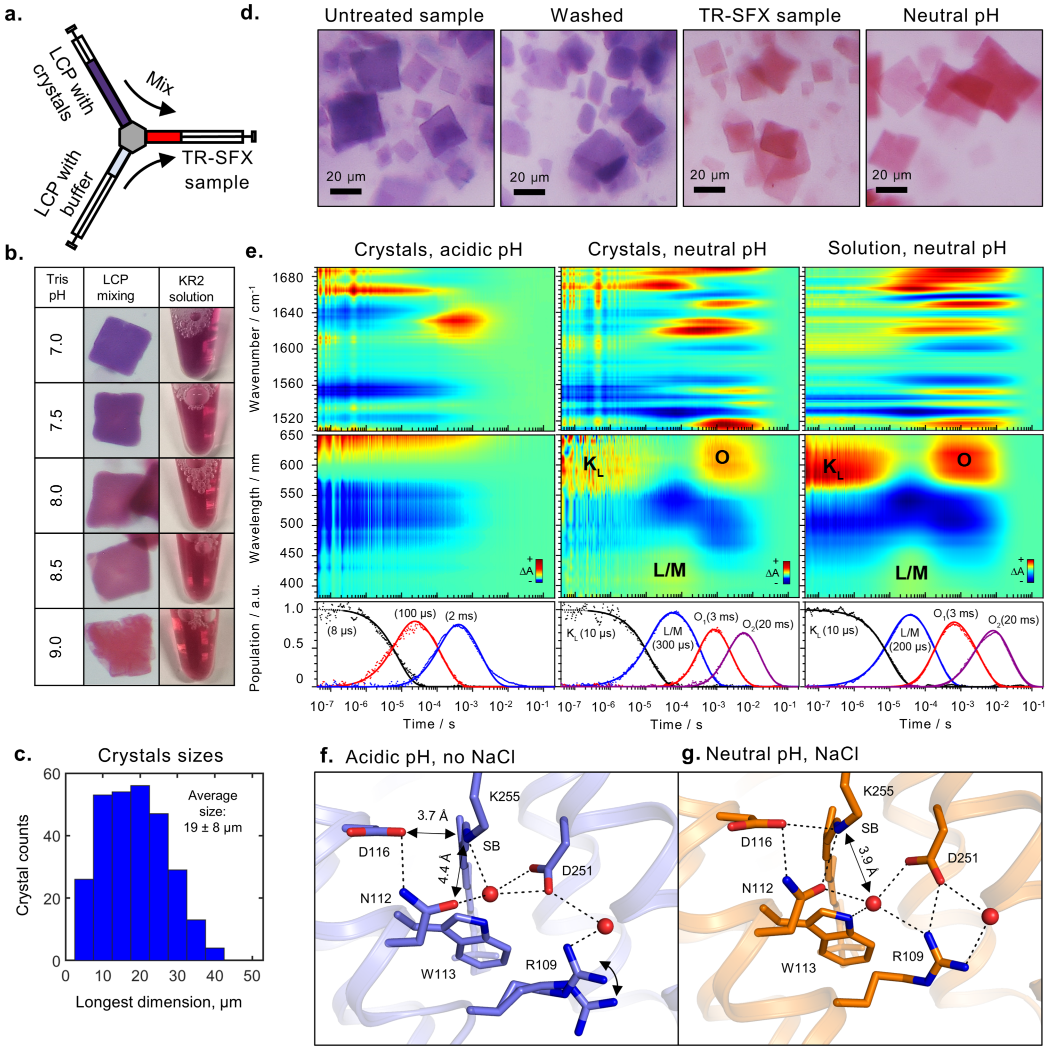
Competing interests The authors declare no competing interests.

Additional information

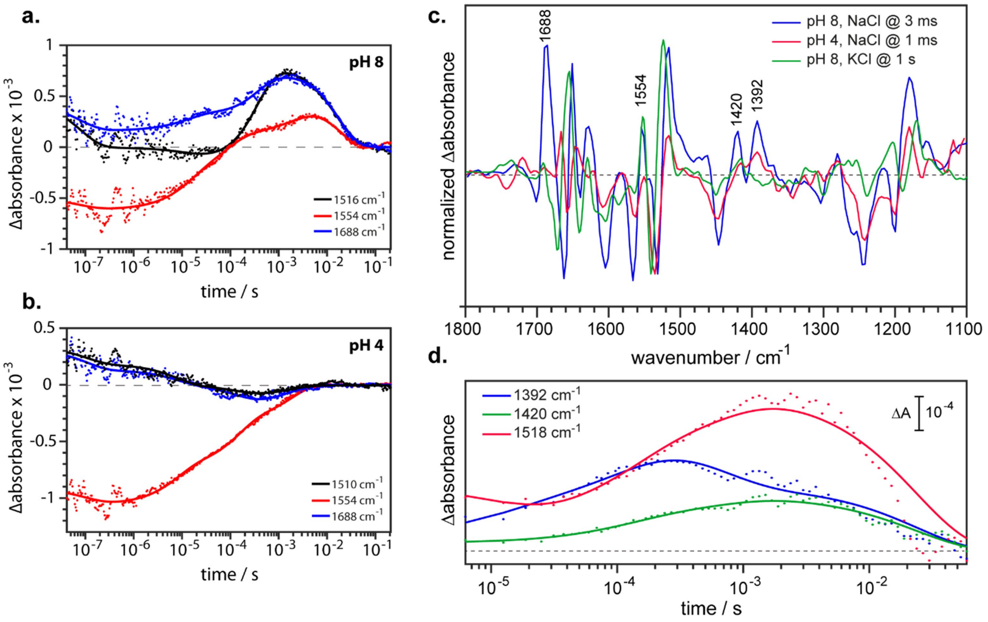
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2307-8.

Correspondence and requests for materials should be addressed to J.S.

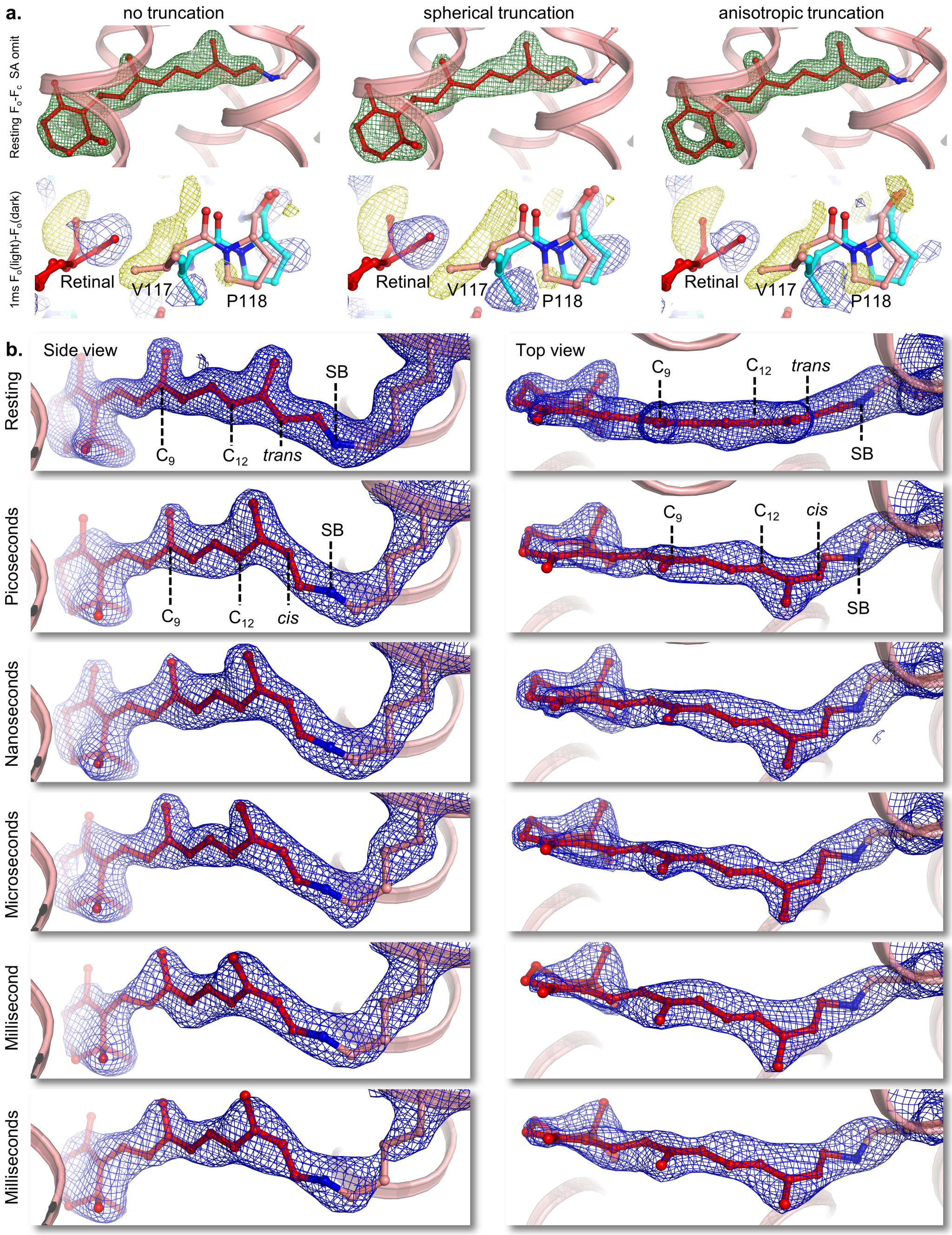
Reprints and permissions information is available at http://www.nature.com/reprints.



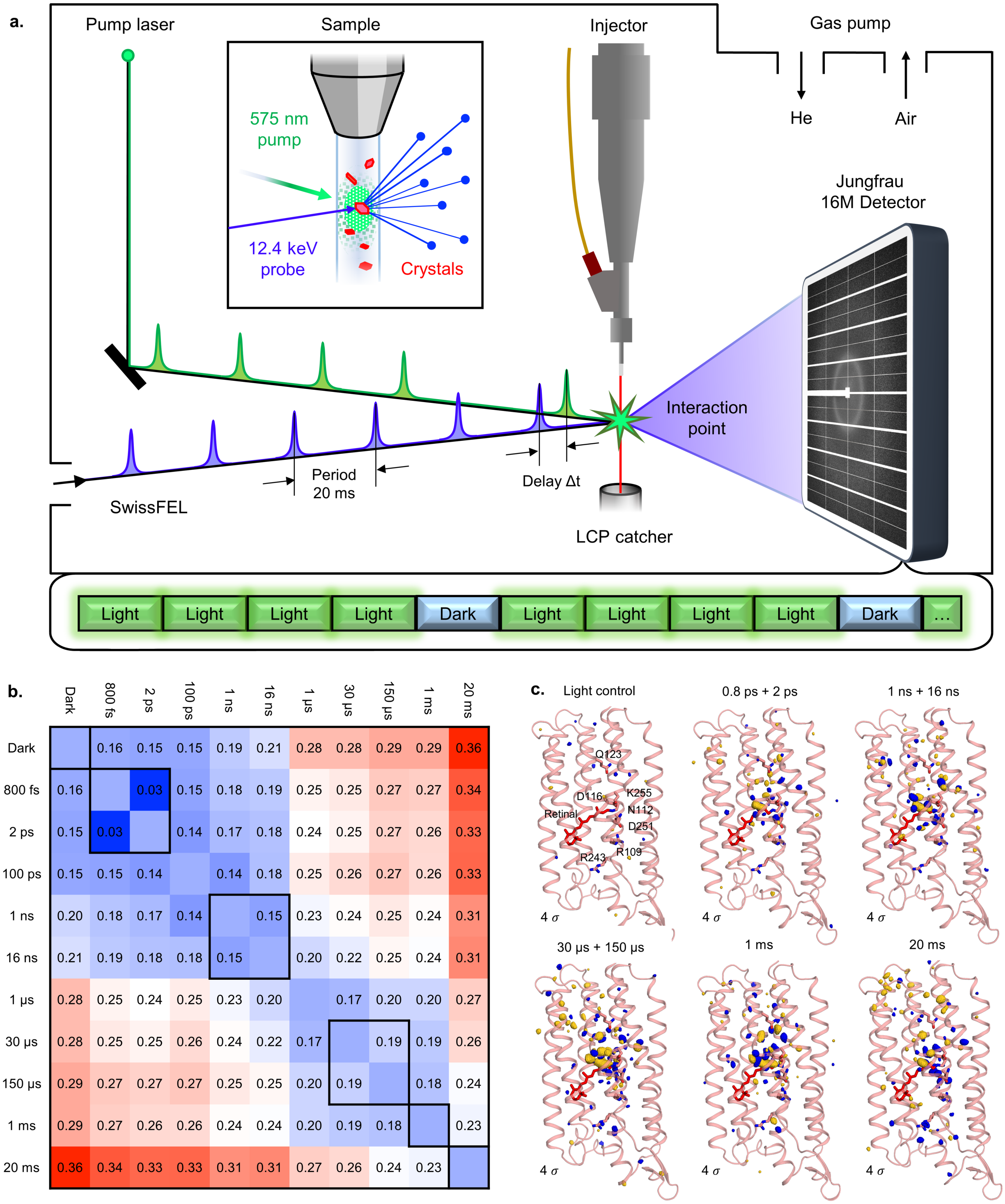
Extended Data Fig. 1 | TR-SFX sample preparation scheme and characterization. a, Two lipidic phases, the first containing KR2 crystals from which the acidic buffer had been washed away and the second containing soaking buffer, were mixed through a three-way coupler[40]. The sodium concentration was adjusted to 150 mM and tests with litmus paper indicate a pH close to 8 in the final mixture. **b**, Left, crystals after TR-SFX sample preparation with varying Tris buffers with pH values from 7.0 to 9.0. Right, KR2 in 200 mM Tris, 150 mM NaCl, 0.02% DDM, 0.004% CHS solution with varied pH of Tris buffer. Note that crystals grown in LCP reach a red colour after mixing with Tris pH 9.0, whereas the solution reaches red colour at pH 8.0. This shift is probably due to residual buffering capacity from the original crystallization conditions, as confirmed by a litmus paper test. **c**, Size distribution of crystals determined by microscopy inspection after TR-SFX sample preparation. **d**, Microscopy picture of KR2 crystals grown at acidic pH and in the absence of NaCl; after washing out the crystallization buffer; the TR-SFX sample prepared via the procedure shown in **a**; and LCP with KR2 crystals soaked directly as control. The colour change upon increasing pH was confirmed in five independent experiments. **e**, Overview of TR-IR traces and visible absorption spectra obtained from blue KR2 crystals in the original acidic crystallization condition, red crystals prepared in analogy to the TR-SFX experiment, and in hydrated film at pH 8 in the presence of sodium chloride (solution). The corresponding lower panels show a kinetic analysis of KR2 intermediates obtained by global fit analysis of the spectroscopic data. Time constants are given in parentheses. Under acidic conditions, KR2 in crystals exhibits an accelerated photocycle. In treated crystals, the critical deprotonation step in the M intermediate occurs with similar kinetics as in purified KR2. **f**, **g**, Detailed view of the retinal binding pocket in the serial crystallographic room temperature structures of KR2 obtained from blue crystals at acidic pH (**f**) and red crystals after soaking in neutral pH and NaCl (**g**). Critical hydrogen bonds with a distance of ≤3.2 Å are indicated by black dotted lines. Arrows signify the distance from the SB to w406 and the D116 counterion. In neutral conditions, w406 has shifted away from the SB while D116 and N112 are now within hydrogen bonding distance. The resulting change in electrostatic environment is responsible for the colour change, as reported previously[2].



Extended Data Fig. 2 | Spectroscopic analysis of sodium binding mode. **a**, **b**, Time-resolved IR absorption changes from KR2 microcrystals at pH 8 (**a**) and at pH 4 (**b**) recorded with tunable quantum cascade lasers as described[54]. The maximum concentration of the O intermediate is reached at around 1–20 ms after pulsed excitation at pH 8 and is characterized by specific marker bands that are absent at pH 4, in particular the C=C stretching vibrational band of retinal at 1,516 cm^{-1} of the O state. The band at 1,688 cm^{-1} has previously been suggested to originate from the C=O stretching mode as a result of sodium binding to an asparagine residue, presumably N112[61]. The band at 1,554 cm^{-1} is tentatively assigned to the asymmetric stretching vibration of a carboxylate that rises upon binding of a sodium ion in bidendate or pseudo-bridging fashion where one of the carboxylate oxygens is interacting with another partner[62,63]. For a detailed analysis of the ligation, the corresponding symmetrical mode needs to be assigned, as the frequency difference between the COO^- asymmetric and symmetric stretch is dependent on the mode of sodium ligation[64,65]. **c**, O(like)-KR2 (ground state) difference spectra recorded under different conditions. Spectra have been scaled to the ground-state bleaching band measured between 1,530 and 1,540 cm^{-1}. It is well-established that KR2 operates exclusively as a sodium pump at neutral pH and in the presence of sodium ions. KR2 acts as proton pump in the presence of KCl but has no (known) function at pH 4[1,28]. It is evident that the band at 1,688 cm^{-1} is most pronounced if sodium is pumped, which supports the assignment to the C=O stretching vibration of N112 upon binding of sodium. **d**, Time traces of the ethylenic stretch of the O state vibrating at 1,518 cm^{-1} and the two candidate vibrational bands at 1,420 cm^{-1} and 1,392 cm^{-1} of the symmetric carboxylate stretching vibration. The band at 1,420 cm^{-1} exhibits similar kinetic behaviour to the one at 1,518 cm^{-1}. Hence, the former vibrational band is tentatively assigned as the symmetrical mode that relates to the asymmetrical vibration at 1,554 cm^{-1} of deprotonated D251 upon ligation of a sodium ion. The difference in frequency between the symmetric and asymmetric modes is with 135 cm^{-1} at the edge of binding in bidentate to pseudobridging mode of model compounds[62,64,65]. This indicates that the two oxygens of D251 are not equidistant from the sodium ion. Such asymmetric ligation is expected in the heterogenous environment of the protein interior as documented by our X-ray structures in the millisecond regime.

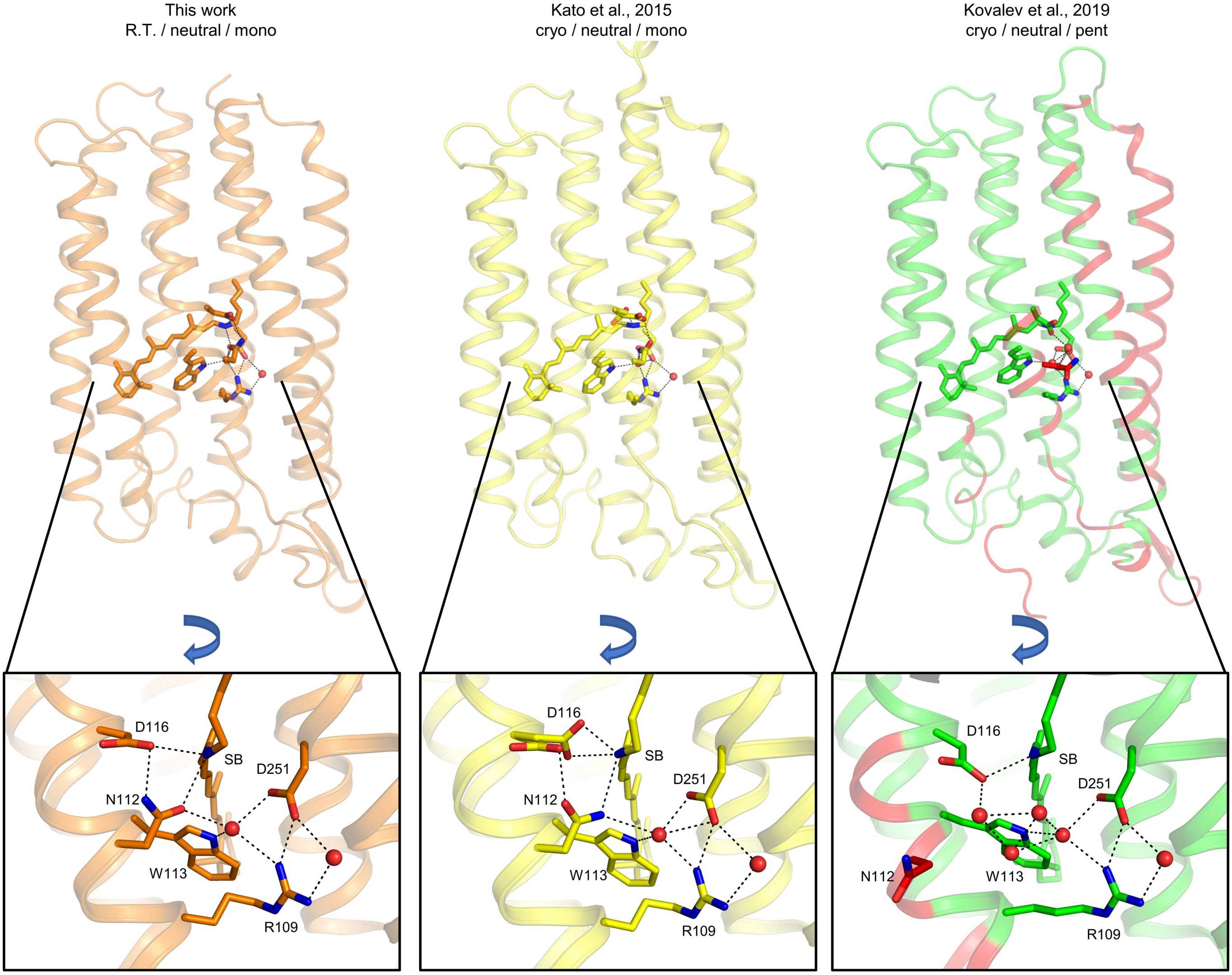


Extended Data Fig. 3 | Comparison of data truncation schemes and changes in retinal over time. a, Top, $F_o - F_c$ simulated annealing omit maps of the resting state at 4σ. No truncation (left) shows the map using all data up to 1.6 Å resolution along a*, b* and c*. Spherical truncation (middle) shows the map using all data up to 2.2 Å resolution along a*, b* and c*. Anisotropic truncation (right) shows the map using data up to 2.3 Å, 2.2 Å and 1.6 Å resolution along a*, b* and c*, respectively, as truncated by the staraniso server[47]. Bottom, F_o(light) − F_o(dark) difference density maps of the 1-ms time delay at the region around retinal and V117 at 3σ. The structure is shown as sticks (salmon, resting state; cyan, 1 ms refined structure). No truncation (left) shows the map using all data up to 1.6 Å resolution. Spherical truncation (middle) shows the map using all data up to 2.2 Å resolution. Anisotropic truncation (right) shows the map using data up to 1.6 Å resolution in c* as truncated by the staraniso server. Overall, the truncated data result in better electron density maps (both for $2F_o - F_c$ maps and F_o(light) − F_o(dark) difference maps), with finer features being resolved. This effect is probably because noise along the missing directions is removed when compared to no truncation, while retaining the high-resolution data along c* when compared to spherical truncation. **b**, The evolution of electron density around the retinal chromophore over time. Retinal and K255 of the refined structures are shown as sticks and the electron density is displayed around them (blue mesh, 2σ). Top, original $2F_o - F_c$ electron density map; panels below show extrapolated $2F_{extra} - F_c$ maps. The extrapolated maps allow us to follow retinal isomerization in detail. In the dark state (top), the middle section or the retinal polyene chain is slightly bent downwards. In the picoseconds range, the isomerization is completed and the polyene appears to be straightened. In our ultrafast data, we did not observe retinal with a pronounced twist in the C13=C14 bond as in bR, with retinal in KR2 reaching a near planar 13-*cis* conformation much earlier along the activation pathway. In the time delays from nanoseconds to milliseconds, the electron density reveals a bend in the retinal molecule resulting from two planes that are twisted against each other. While the exact dihedral angles cannot be refined realistically on the basis of the extrapolated data, the bend seems to originate from the C9=C10−C11=C12 dihedral angle as suggested for the L, M and O intermediates based on time-resolved FTIR[18] and resonance Raman spectroscopy[32]. After 20 ms, a definite conclusion concerning the retinal isomer is difficult. The extrapolated maps suggest that a fraction of the retinal molecules may have already re-isomerized to the all-*trans* conformation, while it is still bent sideways.

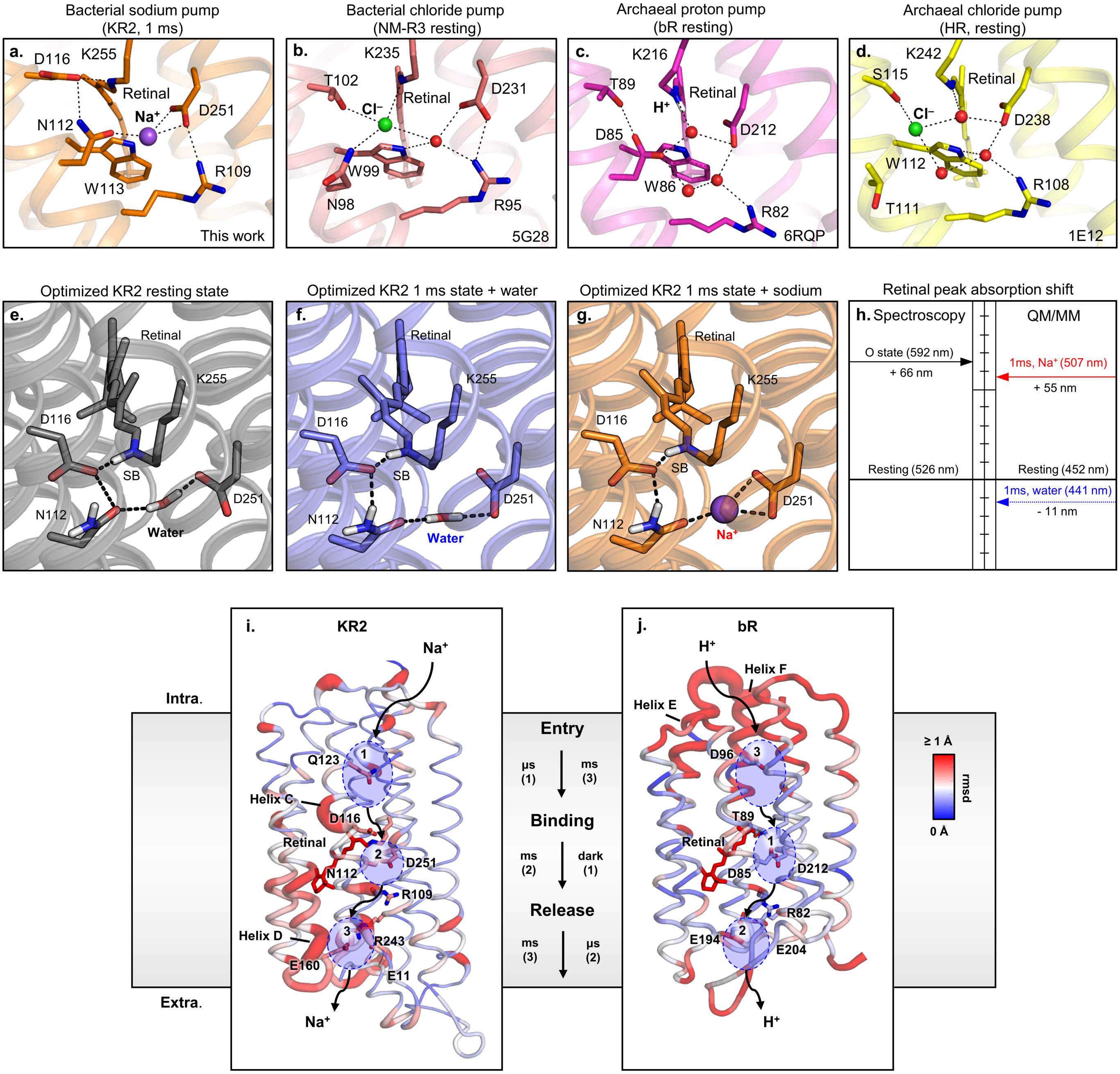


	Dark	800 fs	2 ps	100 ps	1 ns	16 ns	1 µs	30 µs	150 µs	1 ms	20 ms
Dark		0.16	0.15	0.15	0.19	0.21	0.28	0.28	0.29	0.29	0.36
800 fs	0.16		0.03	0.15	0.18	0.19	0.25	0.25	0.27	0.27	0.34
2 ps	0.15	0.03		0.14	0.17	0.18	0.24	0.25	0.27	0.26	0.33
100 ps	0.15	0.15	0.14		0.14	0.18	0.25	0.26	0.27	0.26	0.33
1 ns	0.20	0.18	0.17	0.14		0.15	0.23	0.24	0.25	0.24	0.31
16 ns	0.21	0.19	0.18	0.18	0.15		0.20	0.22	0.25	0.24	0.31
1 µs	0.28	0.25	0.24	0.25	0.23	0.20		0.17	0.20	0.20	0.27
30 µs	0.28	0.25	0.25	0.26	0.24	0.22	0.17		0.19	0.19	0.26
150 µs	0.29	0.27	0.27	0.27	0.25	0.25	0.20	0.19		0.18	0.24
1 ms	0.29	0.27	0.26	0.26	0.24	0.24	0.20	0.19	0.18		0.23
20 ms	0.36	0.34	0.33	0.33	0.31	0.31	0.27	0.26	0.24	0.23	

Extended Data Fig. 4 | Experimental setup and evolution of structural changes over time. a, KR2 crystals were extruded in a stream of LCP from a high-viscosity injector. At the interaction point, the crystals were pumped with 575-nm femtosecond laser pulses before probing for structural changes with near parallel 12-keV XFEL pulses arriving from SwissFEL with a specific time delay Δ*t*. The diffraction patterns were collected with a Jungfrau 16 M detector in a series of four light-activated patterns and one dark pattern, for which the visible pump laser was blocked. To reduce background through diffuse scattering of XFEL radiation in air, the experimental chamber was pumped down to 100–200 mbar while the residual air was replaced by helium. **b**, The evolution of structural changes in KR2 over time can be followed in a matrix of r.m.s.d. between all protein and retinal atoms (total of 2,703) in individual KR2 structural snapshots. The numerical r.m.s.d. values have been determined using the program pymol and are coloured in a blue–white–red gradient. The black boxes highlight time delays at which we combined data based on manual inspection of electron density maps and the evolution of photo intermediates in KR2 crystals determined by TR-IR spectroscopy. The approach was inspired by a recently published tool for visualizing protein motions in time-resolved crystallography[66] but relies on refined atom positions instead of electron density changes. **c**, Difference electron density maps (F_o(light) − F_o(dark), negative density in gold and positive density in blue, contoured at 4*σ* and shown together with the KR2 resting state) obtained at the indicated time delays. The first panel is included as a control and shows difference electron density obtained from 50,000 patterns collected with laser off and 100,000 images from dark patterns obtained using the four light/one dark cycle used during TR-SFX data collection. For orientation, important residues discussed in the main text are shown as sticks.



Extended Data Fig. 5 | Comparison of monomeric and pentameric KR2 structures. Overall the structures at neutral pH reported in this work (left), in Kato et al.[2] (middle) and in Kovalev et al.[20] (right) are very similar (r.m.s.d. of Cα atoms of 0.30 and 0.72 Å, respectively) with the positions of residues and hydrogen bonding pattern (lower insets, black dotted lines, defined as distance ≥3.2 Å) in the binding pocket well preserved. Depending on conditions, the pocket in the pentameric structure can adopt the extended conformation shown. Here N112 is rotated out into the interaction interface between two KR2 protomers (coloured red) and the space is filled with water molecules. As the pentameric resting state can adopt a more open conformation in the SB region[3], we cannot exclude the possibility that the electrostatic gate does not fully open in our crystals formed from monomeric KR2 (see main text). However, an unpublished steady-state structure of pentameric KR2 under continuous illumination (PDB 6XYT, Gordeliy group, IBS Grenoble) shows N112 rotated back into the binding pocket in a conformation very similar to our time-resolved millisecond states. As the retinal is modelled in the *trans* configuration, the 6XYT structure may represent a later intermediate compared to what we observe. The opening along the retinal (calculated by the program Caver[39]) in our 30 µs + 150 µs structure is 1.4 Å, which is very close to the 1.9 Å needed for a sodium ion to pass. Stable structural intermediates that accumulate in time-resolved studies of molecule ensembles do not necessarily reveal all functional states[33] and it seems reasonable that a gate should form only when sodium is in close proximity to the SB. This would allow fast transfer of sodium to the extracellular side upon deprotonation. The gate would then collapse immediately upon sodium binding on the extracellular side, coinciding with reprotonation of the SB. Such a mechanism would efficiently prevent sodium back leakage. Differences in the position of N112 between the monomeric and pentameric resting states could furthermore explain the long-range effects of mutations in the oligomerization interface on the photocycle[17] and sodium pumping efficiency[1,17,20].



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Comparison of ion binding sites in selected members of the rhodopsin family and QM/MM simulations. **a**, Sodium binding site in *Krokinobacter eikastus* rhodopsin 2 (KR2, this work). **b**, Chloride binding site in the related bacterial pump *Nonlabens marinus* rhodopsin 3 (NM-R3, PDB 5G28[67]). **c**, Protonated SB in bacteriorhodopsin (bR, PDB 6RQP[26]). **d**, Chloride binding site in halorhodopsin (HR, PDB 1E12[68]). The retinal chromophore together with selected interactions and amino acid side chains are shown as dotted lines or sticks. Besides overall similarity in ion binding between bacteria and archea, the suggested binding site in KR2 is in line with quantum chemical calculation of absorbance changes upon sodium binding. **e**–**g**, The QM/MM-optimized geometry of the resting state (**e**) and the 1-ms structure with water (**f**) or sodium (**g**) bound to N112, D251. **h**, Comparison of the absorption shifts of the states measured in UV/Vis difference spectroscopy with the calculated QM/MM excitation wavelengths. Absolute values are given in parentheses. **i**, **j**, The extent and location of structural changes in KR2 at 20 ms (**i**) and bR at 15 ms (**j**) after activation (blue-to-red gradient and ribbon width indicate r.m.s.d. to the resting state, bR coordinates taken from previous work[26]). The light-driven pumping against a concentration gradient is achieved with smaller conformational changes in the order of 1–2 Å in sodium pumping KR2 compared to the more elusive protons pumped by bR[24,26]. The translocation path for sodium and protons (arrows), however, is similar for both proteins and includes three critical sites (see reviews on bR structural dynamics[69] and activation mechanism[70]). Close to the water-filled cavity on the cytoplasmic side of KR2, Q123 of the NDQ motif is the most likely location for where the sodium ion loses its water coordination shell. The position is analogous to D96 of the DTD motif in bR, which is the primary donor for reprotonation at the end of the pumping cycle. The second critical site is formed by D116, N112 and D251, which correspond to D85, T89 and D212 in bR and coordinate sodium and proton transport in the SB region. The role of R109 in switching from water to sodium binding in KR2 is analogous in position and function to that of R82, which regulates the transfer of protons towards the release group in bR. The proton release group in bR is formed by E194 and E204, with the position overlapping well with the second sodium binding site between E11, N106 and E160 at the extracellular side of KR2. Beside these similarities, the sequence of events is clearly different between the two outward ion pumps. Light-driven pumping in bR starts with (1) a prebound proton followed by (2) release and (3) reloading. In KR2 the process is shifted in sequence with (1) entry, (2) binding and (3) release with a corresponding adaptation of the photocycle intermediates. The similar ion binding modes between bacteria and archea, together with how retinal isomerization is used to drive them, provides a notable example of the evolutionary economy of nature. The adaptation of this common leitmotif is particularly interesting in the case of KR2, where substrate binding has been shifted from the stable resting state into a transient intermediate.

Extended Data Table 1 | X-ray statistics

	Dark-acidic	Dark-neutral	800 fs	2 ps	100 ps	1 ns	16 ns	1 µs	30 µs	150 µs	1 ms	20 ms	800fs + 2ps	1ns + 16ns	30µs + 150µs
Data collection															
Space group	I222														
Cell dimensions															
a, b, c (Å)	41.5, 84.5, 235.6														
α, β, γ (°)	90, 90, 90														
Indexed patterns	47’518	158’832	47’801	47’595	43’162	45’045	50’165	55’041	50’754	48’383	60’629	48’329	95396	95210	99137
Indexing rate (%)	12.51	11.30	10.68	15.28	9.65	14.11	14.28	11.87	9.00	9.61	12.23	9.16	-	-	-
Overall statistics excluding anisotropic shell 12.4 Å – 2.3 Å (Overall statistics 12.4 Å – 1.60 Å including anisotropic shell)															
Resolution along a*; b*; c* (Å)	2.0; 2.3; 1.6	2.2; 2.3; 1.6	2.1; 2.3; 1.7	2.1; 2.3; 1.7	2.1; 2.20; 1.6	2.1; 2.20; 1.6	2.1; 2.3; 1.7	2.1; 2.3; 1.7	2.1; 2.3; 1.7	2.1; 2.3; 1.7	2.1; 2.2; 1.7	2.1; 2.20; 1.6	2.0; 2.2; 1.6	2.0; 2.2; 1.6	2.0; 2.2; 1.6
No. reflections	18’807 (28’904)	18’873 (31’741)	18’840 (26’799)	18’840 (27’107)	18’872 (29’755)	18’847 (28’057)	18’827 (26’141)	18’828 (26’237)	18’793 (25’531)	18’812 (26’123)	18’847 (27’051)	18’863 (29’034)	18’874 (29’870)	18’872 (30’047)	18’861 (28’402)
Completeness (%)	99.7 (52.3)	100 (57.5)	99.8 (48.5)	99.8 (49.1)	100 (53.9)	99.9 (50.8)	99.8 (47.3)	99.8 (47.5)	99.6 (46.2)	99.7 (47.3)	99.9 (49.0)	99.9 (52.6)	100 (54.1)	100 (54.4)	99.9 (51.4)
Multiplicity	608 (505)	1’640 (1’328)	522 (444)	497 (419)	475 (395)	507 (431)	547 (469)	563 (481)	505 (439)	479 (413)	594 (505)	531 (445)	1018 (834)	1053 (869)	983 (827)
R_{split} (%)	6.1 (6.8)	4.3 (4.9)	8.0 (8.6)	8.3 (9.0)	8.4 (9.1)	7.8 (8.4)	7.3 (7.8)	7.4 (8.0)	7.1 (7.7)	7.8 (8.4)	6.9 (7.5)	7.6 (8.2)	5.7 (6.3)	5.3 (5.9)	5.3 (5.9)
$CC_{1/2}$	1.00 (1.00)	1.00 (1.00)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	0.99 (0.99)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$<I/\sigma(I)>$	11.4 (8.3)	16.5 (11.1)	9.2 (7.1)	9.1 (7.0)	9.2 (6.7)	9.2 (7.0)	9.6 (7.5)	9.7 (7.6)	9.4 (7.5)	9.3 (7.3)	10.4 (7.9)	9.8 (7.2)	12.8 (9.1)	13.1 (9.2)	13.1 (9.6)
High resolution statistics of the anisotropic shell (2.3 Å - 1.60 Å)															
No. reflections	10’097	12’867	7’959	8’267	10’883	9’210	7’314	7’409	6’738	7’311	8’204	10’171	10’996	11’175	9’541
Completeness (%)	27.7	35.4	21.9	22.7	29.9	25.3	20.1	20.4	18.5	20.1	22.6	28.0	30.2	30.7	26.2
Multiplicity	331	872	261	241	256	274	269	272	256	244	302	285	519	559	518
R_{split} (%)	38.3	30.0	43.8	44.3	38.2	41.0	45.2	44.0	45.8	43.8	42.1	38.2	37.0	35.4	37.1
$CC_{1/2}$	0.90	0.95	0.86	0.84	0.89	0.89	0.84	0.86	0.85	0.86	0.87	0.89	0.91	0.92	0.91
$<I/\sigma(I)>$	2.5	3.1	2.2	2.2	2.5	2.4	2.2	2.2	2.2	2.2	2.3	2.5	2.6	2.7	2.6

Extended Data Table 2 | Refinement statistics

	Dark-acidic	Dark-neutral	800fs + 2ps	1ns + 16ns	30µs + 150µs	1ms	20ms
Data collection							
Space group	I222						
Cell dimensions							
a, b, c (Å)	41.5, 84.5, 235.6						
α, β, γ (°)	90, 90, 90						
Resolution (Å)	12.36 – 1.6* (2.3 – 1.6)	12.36 – 1.6* (2.3 – 1.6)	12.36 – 2.25 (2.33 – 2.25)	12.35 – 2.25 (2.33 – 2.25)	12.35 – 2.25 (2.33 – 2.25)	12.36 – 2.5 (2.58 – 2.5)	12.36 – 2.5 (2.58 – 2.5)
R_{split}	6.8 (38.3)	4.9 (30.0)	5.8 (24.29)	5.4 (24.1)	5.4 (27.3)	6.6 (17.3)	7.4 (17.1)
$<I/\sigma(I)>$	8.3 (2.5)	11.1 (3.1)	12.2 (3.5)	12.5 (3.7)	12.5 (3.3)	12.2 (5.1)	11.3 (5.6)
$CC_{1/2}$	1.00 (1.00)	1.00 (1.00)	0.99 (0.95)	0.99 (0.96)	0.99 (0.96)	0.99 (0.97)	0.99 (0.96)
Completeness (%)	52.3 (27.7)*	57.5 (35.4)*	99.9 (98.6)	99.9 (98.6)	99.7 (97.5)	99.9 (100)	99.9 (100)
Multiplicity	505 (331)	1328 (872)	994 (643)	1028 (660)	960 (616)	647 (443)	579 (400)
Refinement							
Resolution (Å)	12.36 – 1.6 (1.64 – 1.6)	12.43 – 1.6 (1.64 – 1.6)	12.36 – 2.25 (2.35 – 2.25)	12.35 – 2.25 (2.35 – 2.25)	12.35 – 2.25 (2.35 – 2.25)	12.36 – 2.5 (2.66 – 2.5)	12.36 – 2.5 (2.66 – 2.5)
No. reflections	28'895 (123)	31'728 (203)	18'021 (1874)	18'073 (1906)	17'917 (1920)	13'446 (2007)	13'268 (1871)
R_{work} / R_{free}	17.67 / 21.3	18.8 / 21.7	27.2 / 32.8	26.5 / 33.2	26.7 / 33.5	25.8 / 32.4	27.8 / 35.3
No. atoms							
Protein	2139	2105	2103	2111	2099	2099	2099
Ligands	242	242	242	242	242	243	243
Water	99	60	53	50	49	49	47
B-factors							
Protein	31.8	36.1	42.1	40.1	33.7	41.7	52.1
Ligands	61.5	62.1	65.6	65.5	53.2	63.6	69.9
Water	46.9	46.1	41.6	43.6	36.3	43.6	50.1
R.m.s. deviations							
Bond lengths (Å)	0.011	0.006	0.008	0.007	0.007	0.007	0.008
Bond angles (°)	1.063	0.735	0.938	0.928	0.928	0.865	0.921
PDB Code	6TK7	6TK6	6TK5	6TK4	6TK3	6TK2	6TK1

* Data are anisotropic, see extended statistics Table 1 and Materials and Methods

Data are anisotropic; see Extended Data Table 1 and Methods.

nature research

Corresponding author(s): Joerg Standfuss

Last updated by author(s): 2020/03/31

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	A full description is provided in the material and methods section.
Data analysis	A full description is provided in the material and methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Resting state coordinates and structure factors have been deposited in the PDB database under accession code 6TK7 (acidic pH) and 6TK6 (neutral pH). Together with the neutral pH resting state structure the light activated data for all time points (800 fs, 2 ps, 100 ps, 1 ns, 16 ns, 1µs, 30 µs, 150 µs, 1ms and 20 ms) was deposited in the mmCIF file. For the refined structures using combined data (800fs+2ps, 1ns+16ns and 30µs+150µs) and single (1ms and 20ms) light activated data sets, coordinates, light amplitudes, dark amplitudes and extrapolated structure factors have been deposited in the PDB database under accession codes 6TK5 (800fs +2ps), 6TK4 (1ns+16ns), 6TK3 (30?s+150?s), 6TK2 (1 ms) and 6TK1 (20 ms).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	No data was excluded.
Replication	Most experimental parts of our study have been replicated repeatedly including expression, purification, crystallization and spectroscopic analysis. All presented crystallographic data were collected in a single X-ray free electron laser beamtime at SwissFEL. However, we were able to collect very similar yet incomplete data in a previous beamtime at the LCLS confirming our results.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Author Correction: New infant cranium from the African Miocene sheds light on ape evolution

https://doi.org/10.1038/s41586-020-2466-7

Correction to: *Nature* https://doi.org/10.1038/nature23456

Isaiah Nengo, Paul Tafforeau, Christopher C. Gilbert, John G. Fleagle, Ellen R. Miller, Craig Feibel, David L. Fox, Josh Feinberg, Kelsey D. Pugh, Camille Berruyer, Sara Mana, Zachary Engle & Fred Spoor

In Extended Data Fig. 3f of this Article, the specimen originally labelled '*P. pygmaeus*' is in fact a specimen of *Pan troglodytes*, and has now been replaced with a confirmed specimen of the Bornean orangutan (*Pongo pygmaeus*). Figure 1 of this Amendment shows the corrected Extended Data Fig. 3. To correct the sample numbers for each species and to add the new *Pongo* specimen now used in Extended Data Fig. 3, the Methods paragraph "The general dental development pattern of KNM-NP 59050… found only in *Hylobates* and *Hoolock*." should read as follows: "The general dental development pattern of KNM-NP 59050, and the advanced I^1 development in particular, were studied in more detail by making comparisons with extant juvenile hominoids and cercopithecoids. These included *Pan troglodytes* (13), *Gorilla gorilla* (3), *Pongo pygmaeus* (2), *Homo sapiens* (6), *Hoolock* sp. (4), *Hylobates muelleri* (1), *Nomascus hainanus* (1); and the cercopithecoids *Papio ursinus* (1), *Cercopithecus petaurista* (1), *Macaca* sp. (2), and *Macaca nigra* (1). These specimens are in the collections of the Musée des Confluences de Lyon and were scanned at the European Synchrotron Radiation Facility, except for one *Pongo* specimen from the Museum of Comparative Zoology, Harvard University that was scanned at the Center for Nanoscale Systems (CNS) at Harvard University, the *Hoolock* material, which is housed in, and was scanned at, the American Museum of Natural History (New York), as well as *Hylobates* and *Nomascus* specimens, which are housed at the Museum für Naturkunde, Berlin, and were scanned at the Max Planck Institute of Evolutionary Anthropology (Leipzig). Results of these comparisons show that the unusual pattern of advanced development of the I^1s is found only in *Hylobates* and *Hoolock*." These errors do not affect any of the observations and conclusions. The original Article has not been corrected.

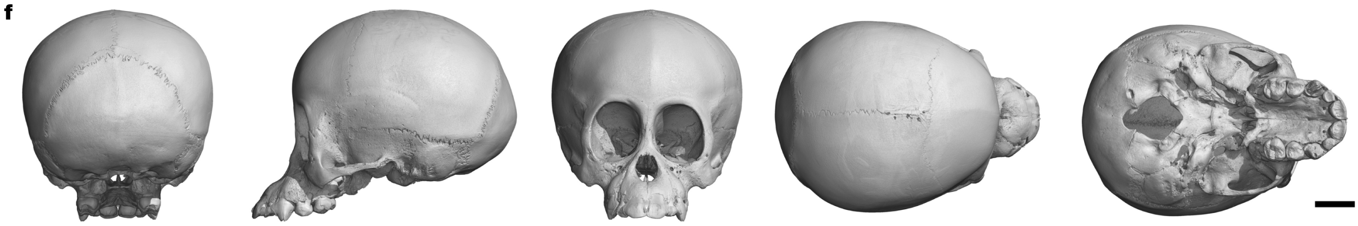


Fig. 1 | This figure shows the corrected Extended Data Fig. 3f. *Pongo pygmaeus*. Scale bar, 2 cm.

Publisher Correction: IGF1R is an entry receptor for respiratory syncytial virus

https://doi.org/10.1038/s41586-020-2437-z

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2369-7

Cameron D. Griffiths, Leanne M. Bilawchuk, John E. McDonough, Kyla C. Jamieson, Farah Elawar, Yuchen Cen, Wenming Duan, Cindy Lin, Haeun Song, Jean-Laurent Casanova, Steven Ogg, Lionel Dylan Jensen, Bernard Thienpont, Anil Kumar, Tom C. Hobman, David Proud, Theo J. Moraes & David J. Marchant

In Fig. 2d of this Article, owing to an error during the production process, the final rightmost *x*-axis label should be 'Ciliated bronchial epithelial cells' and not 'Lung fibroblasts'. This error has been corrected online.

Publisher Correction: Resolving acceleration to very high energies along the jet of Centaurus A

https://doi.org/10.1038/s41586-020-2455-x

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2354-1

The H.E.S.S. Collaboration

In the print version of this Article, M. Holler, M. de Naurois, F. Rieger, D. A. Sanchez and A. M. Taylor should have been indicated as corresponding authors in the author list, all with email address contact. hess@hess-experiment.eu. Additionally, the line 'Correspondence and requests for materials should be addressed to M.H., M.d.N., F.R., D.A.S. or A.M.T.' should have been included at the end of the paper. The Article is correct online.

Matters arising

In vivo quantification of mitochondrial membrane potential

https://doi.org/10.1038/s41586-020-2366-x

Received: 27 November 2019

Accepted: 20 April 2020

Nathaniel M. Alpert[1]✉, Matthieu Pelletier-Galarneau[1,2], Yoann Petibon[1], Marc D. Normandin[1] & Georges El Fakhri[1]

ARISING FROM: M. Momcilovic et al. *Nature* https://doi.org/10.1038/s41586-019-1715-0 (2019)

The recent paper by Momcilovic et al.[1] presents important results on mitochondrial metabolism differences among various lung cancer subtypes in mouse. However, the work[1] propagates critical misunderstandings and omissions about the underlying basis for the application of voltage-sensing tracers. The principle of imaging of mitochondrial membrane potential ($\Delta\Psi_m$) with radiolabelled positron emission tomography (PET) probes is based on benchtop in vitro methods. In vitro quantification of $\Delta\Psi_m$ has been established for decades using various lipophilic cations such as ^{3}H-tetraphenylphosphonium (^{3}H-TPP$^+$)[2–4]. In all cases, the Nernst equation is used to relate the equilibrium concentration of the probe on either side of a membrane to the electric potential across the membrane[5]. Importantly, the Nernst equation is only valid when probe concentrations are at equilibrium—that is, the concentrations on either side of the membrane are not changing with time.

Momcilovic et al. claim to "measure mitochondrial membrane potential in non-small-cell lung cancer in vivo using a voltage-sensitive PET radiotracer"[1], the lipophilic cation 4-[^{18}F]fluorobenzyl-triphenylphosphonium (^{18}F-BnTP). It must be pointed out that, contrary to the authors' claim, they do not measure membrane potential. Their work is based on the empirical endpoint, per cent dose per gram-tissue (normalized by the %dose of the myocardium). %dose of ^{18}F-BnTP is time-dependent and a function of several physiological variables, including the level of ^{18}F-BnTP in plasma, the fractional volumes of extracellular space (ECS), the cellular membrane potential, and also $\Delta\Psi_m$. The need to normalize the endpoint to that of some other organ that may evolve differently in time is also a problem for reproducibility and clinical translation. We have previously studied these crucial aspects of how the research and clinical translation of voltage-sensing compounds can take place[6]. We have shown that a unique absolute endpoint, in units of millivolts (mV), can be obtained during secular equilibrium of a radiotracer such as ^{18}F-TPP$^+$ and ^{18}F-BnTP. Notably, we have shown that, at steady state, the tracer concentration in tissue depends nearly linearly on the tracer concentration in plasma, on the fractional volumes of mitochondria and on the ECS, but depends exponentially on the sum of the cellular and mitochondrial membrane potential:

$$\bar{C}_T \approx \bar{C}_p(1-f_{ECS})f_{mito}\,e^{-\beta(\Delta\Psi_c+\Delta\Psi_m)}, \qquad (1)$$

where $\bar{C}_T$ and $\bar{C}_p$ (in units of Bq ml^{-1}) are, respectively, the steady-state tracer concentration in tissue and plasma, $\Delta\Psi_c$ and $\Delta\Psi_m$ [mV] represent, respectively, the cellular and mitochondrial membrane potentials, f_{ECS} and f_{mito} (unitless) represent, respectively, the volume fractions of ECS and mitochondria, and β [mV^{-1}] is a constant term representing the ratio of known physical constants.

It is also interesting to note that at steady state, equation (1) can be used to express the result in terms of percent dose:

$$\%\bar{D}_T \approx \%\bar{D}_p(1-f_{ECS})f_{mito}\,e^{-\beta(\Delta\Psi_c+\Delta\Psi_m)}, \qquad (2)$$

where we have expressed the result as a ratio of %dose fractions ($\%\bar{D}_{T,p}$). We have previously found that the plasma tracer concentration C_p decreases monotonically after bolus injection of ^{18}F-TPP$^+$, meaning that the %dose index will not be time-invariant, which may limit the reliability of this endpoint and its usefulness in research and clinical translation.

Although Momcilovic et al.[1] did show similar mitochondrial density (f_{mito}) for both cancer subtypes with the pan-mitochondrial marker TOM12, they did not account for other critical variables such as the volume fraction of ECS and the plasma tracer concentrations. Finally, adequate quantification of the membrane potential requires measurements of equilibrium tissue and plasma concentrations[6], which cannot be done with a bolus radiotracer injection and delayed imaging protocol as performed in the study.

Nevertheless, the results of Momcilovic et al.[1] add to the body of evidence supporting the potential role of in vivo assessment of mitochondrial status in oncology. The authors rightly pointed out that imaging membrane potential might represent a valuable resource for the evaluation of mitochondrial activity in several areas of research including ageing, physiology, and diseases. However, for successful translation of the methodology to human research and ultimately to the clinic, accurate and reproducible quantification is necessary. This can only be achieved with proper techniques and accounting for all critical variables.

1. Momcilovic, M. et al. In vivo imaging of mitochondrial membrane potential in non-small-cell lung cancer. *Nature* **575**, 380–384 (2019).
2. Kauppinen, R. Proton electrochemical potential of the inner mitochondrial membrane in isolated perfused rat hearts, as measured by exogenous probes. *Biochim. Biophys. Acta* **725**, 131–137 (1983).
3. Rottenberg, H. Membrane potential and surface potential in mitochondria: uptake and binding of lipophilic cations. *J. Membr. Biol.* **81**, 127–138 (1984).
4. Wan, B. et al. Effects of cardiac work on electrical potential gradient across mitochondrial membrane in perfused rat hearts. *Am. J. Physiol.* **265**, H453–H460 (1993).
5. Nernst, W. Die elektromotorische Wirksamkeit der Jonen. *Z. Phys. Chem.* **4**, 129–181 (1889).
6. Alpert, N. M. et al. Quantitative in vivo mapping of myocardial mitochondrial membrane potential. *PLoS One* **13**, e0190968 (2018); correction **13**, e0192876 (2018).

Acknowledgements The authors gratefully acknowledge support from NIH grants P41-EB022544 and R01-HL137230. These grants were awarded as part of the standard peer review process for new grants. The study design and analysis are solely the work of the study authors and the funders had no role in the study.

---

[1]Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Medical Imaging, Montreal Heart Institute, Montreal, Quebec, Canada. ✉e-mail: alpert@pet.mgh.harvard.edu

Author contributions N.M.A. contributed the overall measurement strategy. M.P.-G. wrote the first draft. Y.P., M.D.N. and G.E.F. edited and revised the manuscript.

Competing interests The authors are employed by The General Hospital Corporation, Partners Healthcare, Boston, MA, USA. US patent US20190125281A1, 'System and method for quantitatively mapping mitochondrial membrane potential' (status pending) is assigned to The General Hospital Corporation and lists N.M.A., G.E.F., M.D.N. and N. Guehl as inventors.

Additional information

Correspondence and requests for materials should be addressed to N.M.A.

Reply to: In vivo quantification of mitochondrial membrane potential

https://doi.org/10.1038/s41586-020-2367-9

Milica Momcilovic[1], Orian Shirihai[4,5], Michael P. Murphy[6,7], Carla M. Koehler[3,5], Saman Sadeghi[2] & David B. Shackelford[1,5]✉

REPLYING TO: N. M. Alpert et al. *Nature* https://doi.org/10.1038/s41586-020-2366-x (2020)

We thank the authors of the accompanying Comment[1] for their interest in our work and acknowledging the biomedical significance of our recent multi-disciplinary study using positron emission tomography (PET) imaging to assess mitochondrial membrane potential in vivo within lung tumours[2]. To accomplish this work we used the ^{18}F-labelled lipophilic cation tracer ^{18}F-BnTP as a tool to image lung tumours in mouse models and identified conserved differences in mitochondrial function within lung tumour subtypes. Our work, along with studies by others in the field[3,4], builds upon prior work by Alpert and colleagues[5] and demonstrates that ^{18}F-BnTP accumulation within tissues is responsive to the voltage across the mitochondrial inner membrane, enabling measurement of relative changes in mitochondrial membrane potential in tissues such as the heart. However, although we acknowledge the work performed by Alpert et al. on measuring mitochondrial membrane potential, we disagree with their claims that our study "represents critical misunderstandings and omissions"[1] regarding the underlying basis of the application of voltage-sensing tracers.

It has been well established that the uptake of the ^{18}F-BnTP tracer into tissues increases with the voltage across the mitochondrial inner membrane and the plasma membrane. However, relating this uptake to voltage is challenging—therefore we imaged mice at 60 minutes post ^{18}F-BnTP injection to ensure that each animal was as close to steady state as possible. We chose the heart as a tissue to normalize uptake of the probe because this tissue reaches a steady-state level of accumulation in around 10 min and shows little change in probe retention over the following 50 min. This enabled us to clearly show changes in probe uptake within the tumours. Although taking multiple plasma samples and kinetic imaging analysis would be interesting, it would add limited value and make the scope of studies performed impractical. Further computed tomography (CT) imaging for measurement of fractional volumes of extracellular space and the plasma tracer concentrations studies to better quantify membrane potential—although appealing—are not practical, and would not translate to clinical situations. Although these extra measurements would partially facilitate absolute quantification, this is not necessary to show the clinically significant relative changes in membrane potential between tumour and control tissue.

The remaining debate lies in our description of the quantitative measurement of mitochondrial membrane potential. Alpert et al.[1] are right to state that we did not directly measure mitochondrial membrane potential in millivolts (mV) and that it was imprecise to use this term in our original paper. Instead, we measured the relative uptake of ^{18}F-BnTP in lung tumours normalized to uptake of the tracer in the heart. The changes in uptake of ^{18}F-BnTP that were measured are reflective of changes in mitochondrial and plasma membrane potentials following the bolus injection of the PET ligand at a set time. Importantly, by carefully measuring the relative uptake of ^{18}F-BnTP in the tumour and the heart tissue at the same dose and time after injection of PET ligand, and using built-in controls that include treating mice with bona fide respiratory-chain inhibitors, we can infer relative differences within each tumour that are consistent with changes in mitochondrial polarization. Our methodology is analogous to in vitro cell culture assays that measure relative changes in mitochondrial membrane potential using dyes, such as TMRE and Rhodamine 123, that accumulate within the mitochondria in a voltage-dependent manner. In these assays, absolute values of the mitochondrial membrane potential in mV are difficult to infer from the fluorescent intensity measurements, because the application of the Nernst equation requires the measurement of many other cell parameters, which are challenging to determine. Instead, the relative differences in membrane potential are presented in arbitrary units (a.u.). The change in units—whether in mV or a.u.—is reflective of differences in the mitochondrial membrane potential. Applying this robust approach to our in vivo work, we have not presented absolute measurements in mV of mitochondrial membrane potential in lung tumours, but instead we have measured the ratio of tracer uptake in lung tumours to that in heart tissue. In essence, the matter that has arisen can be reduced to a difference in methodology—we measured and quantified a ratio instead of an absolute value in mV. Using this methodology we are able to accurately deduce changes in ^{18}F-BnTP uptake in lung tumours that reflect perturbations in the mitochondrial membrane potential.

We have demonstrated that our approach is reproducible in mice and shown that ^{18}F-BnTP is a valuable tool for the in vivo study of mitochondrial biology in mouse models of lung cancer. We recognize that clinical translation of ^{18}F-BnTP into humans will require additional studies. However, this does not detract from the impact or accuracy of pre-clinical studies, or the importance of this probe in future clinical applications. We agree that clinical translation of probes such as ^{18}F-BnTP will require consideration of other variables, rigorous feasibility studies and cooperation between multiple disciplines. Finally, our work[2] as well as that of Alpert and colleagues[5] is critical for advancing technologies to characterize mitochondrial function in vivo, and as such, we are thankful to the authors of the accompanying Comment[1] for addressing this matter in open and respectful debate.

1. Alpert, N. M., Pelletier-Galarneau, M., Petibon, Y., Normandin, M. D. & El Fakhri, G. In vivo quantification of mitochondrial membrane potential. *Nature* https://doi.org/10.1038/s41586-020-2366-x (2020).

---

[1]Division of Pulmonary and Critical Care Medicine, Department of Medicine, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [2]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [3]Department of Chemistry and Biochemistry, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [4]Department of Endocrinology, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [5]Jonsson Comprehensive Cancer Center, David Geffen School of Medicine at the University of California, Los Angeles, CA, USA. [6]MRC Mitochondrial Biology Unit, University of Cambridge, Cambridge, UK. [7]Department of Medicine, University of Cambridge, Cambridge, UK. ✉e-mail: dshackelford@mednet.ucla.edu

2. Momcilovic, M. et al. In vivo imaging of mitochondrial membrane potential in non-small-cell lung cancer. *Nature* **575**, 380–384 (2019).
3. Madar, I. et al. Characterization of membrane potential-dependent uptake of the novel PET tracer ^{18}F-fluorobenzyl triphenylphosphonium cation. *Eur. J. Nucl. Med. Mol. Imaging* **34**, 2057–2065 (2007).
4. Madar, I. et al. Characterization of uptake of the new PET imaging compound ^{18}F-fluorobenzyl triphenyl phosphonium in dog myocardium. *J. Nucl. Med*. **47**, 1359–1366 (2006).
5. Alpert, N. M. et al. Quantitative in vivo mapping of myocardial mitochondrial membrane potential. *PLoS One* **13**, e0190968 (2018).

Acknowledgements D.B.S. was supported by UCLA Jonsson Comprehensive Cancer Center grant P30 CA016042, Department of Defense LCRP grant W81XWH-13-1-0459 and NIH/NCI R01 CA208642-01. M.M. was supported by American Cancer Society Research Scholar grant RSG-16-234-01-TBG.

Author contributions This Reply was submitted by D.B.S. and senior authors of the original Article, as well as M.P.M., all of whom have expertise directly related to the topic discussed in the accompanying Comment. D.B.S. wrote this Reply with expertise and critical feedback contributed by M.M., S.S., O.S., C.M.K. and M.P.M.

Competing interests D.B.S., M.M. and S.S. have filed a provisional patent for the use of ^{18}F-BnTP to guide the use of complex I inhibitors in the treatment of lung cancer.

Additional information

Correspondence and requests for materials should be addressed to D.S.

Reprints and permissions information is available at http://www.nature.com/reprints.

Work

Your story

Send your careers story to: naturecareerseditor @nature.com

ROMY ARROYO FERNANDEZ/NURPHOTO/GETTY

Campaigners march through Amsterdam on 1 June to protest against anti-Black violence in Europe and the United States.

FIGHTING RACISM DEMANDS MORE THAN JUST WORDS

Frustrated and exhausted by systemic bias in the science community, Black researchers call on their colleagues and institutions to take action.

Black academics are calling out racism in science, recounting behaviours ranging from overt acts to microaggressions, using social-media hashtags such as #BlackInTheIvory.

A study in April (B. Hofstra *et al. Proc. Natl Acad. USA* **117**, 9284–9291; 2020) highlighted how students from under-represented groups innovate more than their white male counterparts do — but receive few to no career benefits from their discoveries, because their contributions are often overlooked. *Nature* spoke to six Black academic researchers about the effects of racism on their careers, their advice to their white colleagues and their thoughts on meaningful institutional actions.

VASHAN WRIGHT

WHITE COLLEAGUES HAVE THE POWER TO CHANGE THE SYSTEM

Before I started out as an undergraduate, I thought a university campus was a magical place. I thought I'd be treated not on the basis of my skin colour, but according to how smart I was and my efforts to make the world a better place. This view changed when I saw a Confederate flag flying on my campus — the first time I'd seen one in real life. It changed when someone used a racial epithet against me and threatened to kill me.

Black academic success happens despite systemic racism and bias. It's hard to tell whether dismissive body language is bias. It's hard to tell why my first paper was in limbo for ten months and rejected without review. It's hard to tell whether people's faces are expressing surprise at my ideas, or at the fact that those ideas are coming from someone who is Black. It's those quiet things that eat away at you.

In March, I started as a postdoc at Woods Hole Oceanographic Institution (WHOI) in Massachusetts, which I had visited a couple of times as a PhD student. But I haven't physically been there in my new position because of the coronavirus. I enjoyed earning a PhD in a laboratory with a diverse range of students at Southern Methodist University in Dallas, Texas, and I'll miss that environment, but I appreciate that WHOI has formed a committee for diversity and inclusion and a partnership-education programme to foster a culture of inclusivity.

More recently, Hendratta Ali, a geoscientist at Fort Hays State University in Kansas, and colleagues crafted an anti-racist action plan for geoscience societies. It calls for them to collect scientifically valid data on diversity, equity and inclusion, to publish accountability reports and to ensure that diversity and racial justice are discussed at well-attended events.

Still, the geosciences are among the least diverse disciplines in science, technology, engineering and mathematics. I sometimes worry that I'll never be able to recreate the diverse environment that I enjoyed during my PhD, but I am committed to that vision as I advance in my career.

Some white faculty members don't want to acknowledge that Black students experience racism. They don't necessarily deny your experiences, but they often look for another explanation to try to protect you. A faculty member of colour knows that they cannot protect you from it.

Scientific communities need to decide where, and for what, they stand by asking their members: what message are you sending if you are not actively being anti-racist and trying to change the system? I say to my white colleagues: you have the most power in the geosciences; you benefit the most from racism and lack of diversity. It is therefore your job to fix the system. But I'll help you.

Vashan Wright is a geophysicist at Woods Hole Oceanographic Institution, Massachusetts.

MARK RICHARDS

COMMIT TO BOLD HIRING TARGETS

It's quite easy for a Black person in the United Kingdom to relate to underlying issues of racism in the United States. Heavy-handed policing is not new to our Black community. Like most people around the world, we are experiencing sadness, anger and frustration. Being one of the few Black academics in my field and at this university, I feel a level of responsibility to do what I can to nurture and inspire the next generation. To be a good ally, it's not quite enough to be neutral. You have to be anti-racist when necessary. Racism is like a virus — attitudes spread when they are validated. To kill the racism virus, you have to distance yourself from that ideology so that eventually there is nowhere for it to be transmitted. The best way to combat white supremacy is to focus on Black excellence.

Until recently, many institutions didn't feel there was a problem. They lamented that no Black students were applying, but a 'What can we do?' attitude prevailed for many years. I think that institutions must look at their data to shed light on whether they are being institutionally racist. If a university is located in the centre of a huge city with a high Black population living locally, but has only a minuscule number of Black students, that has to tell you something.

Physicist Mark Richards wants institutions to take on more students from minority groups.

MARK RICHARDS/IMPERIAL COLLEGE

At Imperial College London, we have been trying for 15 years to address a shortage of Black students and those from other under-represented groups, through an advisory group called Imperial as One. Yet representation of people from all minority ethnic groups remains low here, at less than 10%. The problem will not be solved simply by employing more faculty members of colour. We have to actually attract students from under-represented groups. Imperial did something decisive and made a commitment to double its intake of students from under-represented groups over the next five years. The push to achieve this goal is due to start this autumn. I hope other academic institutions will replicate these types of action.

Beyond simply treating all people as equals, enough data exists to show that diverse teams are more productive. Posing the question, 'Is your university being as diverse as it could be?' is essentially the same as asking, 'Is the university being as productive as it could be?' What leader doesn't want to ask that question?

Mark Richards is a physicist at Imperial College London.

KISHANA TAYLOR

CONSIDER 'CLUSTER HIRING'

I met my postdoc adviser, virologist Sam Díaz-Muñoz, in 2017 at a Twitter meet-up, at a conference organized by the American Society for Microbiology in New Orleans, Louisiana. At the end of the conversation, he asked me to apply for a job in his new lab at the University of California, Davis.

My PhD experience, at the University of Georgia in Athens, had dealt a blow to my self-esteem. I was the only Black graduate student in a department with no Black professors. I was by myself, in terms of representation, in a state where Black people make up 30% of the population. When I met my postdoc adviser, I hadn't yet published anything and had no confidence that I would be competitive. Had he not contacted me a second time, I would not have pursued that position.

During my PhD programme, between 2013 and 2018, high-profile shootings of unarmed Black people — notably Eric Garner, Tamir Rice, who was just 12 years old, and Michael Brown — prompted a number of criminal trials, but


DOANISE THOMPSON

Virologist Kishana Taylor calls for more scientists of colour in leadership positions.

none secured a conviction. I'd come into the lab feeling heavy and upset after these deaths. No one noticed. I tried to talk about my feelings to a staff member, who told me not to get too upset, because we didn't know what had happened. I learnt quickly to not even bother trying to have those conversations.

I was going to leave science after my PhD unless I found a lab that valued diversity. Although I found a principal investigator who does, my department at Davis doesn't have any Black professors. Academic institutions need to take action instead of just saying they value students from under-represented minority groups. We need more scientists of colour in leadership positions.

Practices such as 'cluster hiring', which can be used to diversify faculties, have been tried at a number of universities and should get more attention. With cluster hiring, universities advertise multiple faculty positions at once, but don't always stipulate specific fields; this can improve the odds that candidates from under-represented minority groups will be selected, making it easier for them to fit in if they are. It's infinitely more lonely, and harder to adjust, if someone is the only Black person to be hired in a department. For example, the entire University of Maryland system has a US National Science Foundation grant to create a pre-professoriate pipeline, hiring transitional postdocs with the intent of bringing them in as faculty members in a year or two.

It's important for white colleagues to ask themselves whether they'd be comfortable walking into a room full of Black or Latin American people — and then to imagine what it's like for us when, every day of our lives, we have to enter rooms full of only white people. My biggest pet peeve is when white colleagues, who do research for a living, ask me for advice on how to be an ally without having done any research. It's not hard to find journal articles that detail the impact of diversity, equity and inclusivity initiatives.

Kishana Taylor is a virologist at the University of California, Davis.

NIKEA PITTMAN
CREATE OPPORTUNITIES FOR DIFFICULT CONVERSATIONS

My university and its diversity and inclusivity committee have released messages of support for the Black community following the protests in the wake of George Floyd's death, but it can feel like there's a lot of silence beyond those messages. It's hard for a Black person to initiate conversations around this topic because of a fear of coming across as aggressive, especially when those conversations might not be welcome. In early June, two white male colleagues (not in my lab) discussed the protests, standing less than a metre away from me, while I pretended not to be hurt by the exclusion.

Almost two weeks later, one of the youngest graduate students in my lab suggested we have a group conversation about race. I intentionally didn't open my mouth until everyone else had spoken. I hadn't heard what my colleagues think about these issues, and it was easier for me to open up once it became apparent that everyone wanted to know how to be an ally to the Black community. If your lab hasn't had this conversation yet, you can still have it — and it can make a big difference.

I've been waiting for more than a decade to hear non-Black people be outraged at the way Black people are treated in the United States, and was hopeful that we could collectively mourn. Even though it's not my responsibility, I want to help make these conversations easier. But I don't always know how to proceed. I jumped on a bandwagon of Black academics on social media who are volunteering to support younger colleagues who feel alone. We need collaboration in academia to tackle these problems. Black scientists can't carry the burden on their own.

Universities can help faculty members to learn how to steer difficult conversations, and to acknowledge the emotional burden of systemic racism. Department heads can encourage their faculty members to initiate and maintain these discussions with their teams.

At the institutional level, so much of the diversity conversation right now is focused on recruiting more students of colour and finding ways to support them once they arrive. And that's important, but it's a very different question from the one that the Black community is struggling with right now: we're asking what we can do to prevent the next instance of police brutality. I saw that the University of Minnesota in Minneapolis had severed a contract with the city's police department, which had provided patrols at large campus events, because of concerns about violence. Those are bold, immediate actions that get to the root of the problem.

On social media, I saw medical students at the University of Washington in Seattle initiate an anti-racism summer reading programme. I really hope that tenured white faculty members will do the research, too, and be able to say: "I've learnt how racism, discrimination and implicit bias affect my Black colleagues." And then they will realize at the next faculty meeting that they can start conversations that their Black colleagues cannot start without putting their careers on the line.

Nikea Pittman is a structural biologist at the University of North Carolina, Chapel Hill.

HENRY HENDERSON
CREATE A WELCOMING ENVIRONMENT

I trained at Tuskegee University in Alabama, a historically Black university, where everyone looked like me. But I started to question whether I wanted to continue in academia when I went to national conferences and saw so few people of colour. Yet I realized that, if I stepped out, it would decrease the chances for someone else who looked like me to climb the academic ladder.

At Vanderbilt University in Nashville, Tennessee, I am one of just two Black postdocs on a floor with at least ten labs. It reminds me there's still a lot to do. We have programmes to support diversity and inclusion, but that's not enough. I've experienced microaggressions. At a poster presentation at a large meeting, every question I got asked was, "Who did all this work?", even though I was first author. It's not unusual for white colleagues to insinuate that people of colour are in certain programmes only because their research is being supported by supplementary grants for members of minority groups. If you create opportunities, honour the policy in place to increase diversity.

When I come to the lab in the evening, I am usually followed by campus police from the parking complex to the building, and they ask to see my badge but let white people go right by. I'm used to that from city police. My mother wants me to wear my ID badge everywhere I go, in the hope that it will keep me safe.

I had considered leaving academia several times. My current principal investigator, oncologist Christine Lovly, saved me. She strives to foster opportunities, and puts me in leadership roles – inviting me, for example, into a group focused on health disparities in cancer – so that I can have a voice and gain experience promoting dialogue in areas relevant to my research. This approach took me by surprise at first. I wondered whether she was overworking me, but she wanted to increase my impact in spaces that I otherwise wouldn't be in. She's encouraged me to apply for a US National Institutes of Health grant, even though I was hesitant because of the gap between success rates for white and African American applicants. I wouldn't be on Twitter if it weren't for her. She saw that I wanted to diversify science, and created this postdoc opportunity so that I could reach back and pull others up, too.

Creating a welcoming environment is very different from inviting Black people to university functions, academic programmes and professional society meetings simply to lift our numbers. We can tell when we're here to be here as opposed to being tokenized. In a welcoming space, we are asked to speak at talks, to offer input, to collaborate and to lead projects.

I worry that there are periods during which society is outraged by racism, only for everything to go back to the way it was within a couple of months. We need a continuous effort to improve racial diversity in science and medicine. Institutions should create more programmes that emphasize not only diversity, but also the retention of Black students in the scientific pipeline, including support throughout their education. I have six nephews who are all interested in science because of me. Imagine if there were more Black academics. How many other kids would be inspired?

Henry Henderson is a cancer biologist at Vanderbilt University Medical Center, Nashville, Tennessee.

ABDULHAKIM ABDI
MAKE HIRING FOR LEADERSHIP POSTS MORE TRANSPARENT

I came to Sweden in 2012 to get a PhD at Lund. Before that, I worked as a geographic-information systems specialist at the Lamont–Doherty Earth Observatory in New York City. There, Robin Bell, head of the research group and current president of the American Geophysical Union, created an environment where everybody's voice was heard and respected. That encouraged me to pursue a PhD and to ask my own research questions.

At the same time, I experienced overt racism in the United States while exploring the outdoors as a birdwatcher. I grew up in Abu Dhabi and started observing birds when I was 14. After I moved to the United States at the age of 21, for my undergraduate degree, I had to explain to police at least half a dozen times that I was simply outside watching birds. The police weren't aggressive, but they told me that I was making people uncomfortable and that I had to move along. So, for the rest of the 11 years that I was in the United States, I stopped birdwatching. Here in Sweden, people might give me strange looks when I'm outdoors, but they don't say anything or call the police. In everyday encounters here, it can be difficult to discern whether people are being racist or whether they are simply introverted culturally.

The student body is pretty diverse in universities across Sweden, but less so in the upper echelons. I'm worried about the lack of transparency in hiring and promotion decisions. What goes on behind closed doors? These problems were detailed in a 2018 report on the prevalence of nepotism in Swedish academia (see go.nature.com/3ip75Yh).

As the report suggested, a combination of increased funding, to help alleviate universities' reliance on hiring 'safe bets', and greater transparency in hiring should reduce nepotism and hopefully eliminate it. Swedish institutions have strong gender-equality programmes, as they should. But there are no programmes that open the higher echelons of academia to qualified and capable researchers from the country's recognized minority groups; these include the Sami and the Roma, as well as other established communities such as people from Iraq and Somalia.

Still, when I was offered a position in the United States soon after I earned my PhD, I decided to stay in Sweden with my family – it made more sense financially because we have subsidized day care and health care.

I think every university should create an ecosystem where Black scientists and those from other minority groups feel comfortable. Academics need to let go of the myth that they have no biases. They need to respect Black people's opinions and invite Black and minority-ethnic academics to be co-authors of papers. And they need to expand their networks to create opportunities for colleagues from under-represented groups.

Abdulhakim Abdi is a physical geographer at Lund University, Sweden.

Physical geographer Abdulhakim Abdi calls for a sustained effort to improve racial diversity.

HATIM ELKHIDIR

Interviews by Virginia Gewin. Interviews have been edited for clarity and length.

Where I work
Zach Ousnamer

Photographed for *Nature* by Rocco Ceselin.

They call me a mechatronics engineer – a hybrid between a roboticist and a mechanical engineer. In the Spacecraft Assembly Facility at the Jet Propulsion Laboratory (JPL) in Pasadena, California, we build and test spacecraft and rovers, the vehicles that can traverse the surface of a destination.

In February, we finished the main assembly of the Mars 2020 rover, Perseverance – due to launch this month or in August – whose mission is to seek signs of ancient life. What's exciting about this rover is that it can take samples of Martian soil, analyse them and seal them in containers for a future mission to collect and bring back to Earth. The rover's complex machinery can drill down to extract core samples 4–5 centimetres long, then bring them inside the rover to process and photograph.

On any mission, we try to ensure that no volatile chemicals, plastics or paints are deposited on sensitive surfaces such as optical lenses. We also take care to avoid contaminating the Mars surface with Earth organisms such as bacteria or spores, which could interfere with getting accurate results. We work in the giant 'clean room' seen here. Before entering, we put on full protective gear: little has changed at the JPL in that respect since the COVID-19 lockdown began.

With Mars 2020, we have to be even more careful because some of the samples collected will come back. When we're working with equipment that will process soil cores, we wear an extra set of single-use, sterile gloves, and a sterile lab gown and goggles. A buddy hands me tools, anything I need, so that I don't contaminate my gloves.

Since February, I've been based at the Kennedy Space Center in Cape Canaveral, Florida, helping to put the finishing touches to the rover.

As I ungown for the day, it's nice to take a step back and think how we've sent rovers to other planets only a handful of times. It's pretty monumental.

Zach Ousnamer is the integration and test engineer for NASA's Mars 2020 mission at the Jet Propulsion Laboratory in Pasadena, California. **Interview by Amber Dance**.